ZOLD: A ZEROTH-ORDER LANGEVIN DYNAMICS FRAMEWORK FOR LLM FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Zeroth-order (ZO) optimization has emerged as a memory-efficient paradigm for fine-tuning LLMs, estimating gradients from forward evaluations alone. However, existing approaches refine the ZO gradient estimator within an SGD-style update and thereby inherit a structural limitation of SGD: the absence of a mechanism for global exploration of the loss landscape. In this paper, we propose a novel framework, Zeroth-Order Langevin Dynamics (ZOLD), which can be viewed as a zeroth-order analogue of stochastic gradient Langevin dynamics. A precise coupling between the smoothing radius and the inverse temperature ties the ZOLD dynamics to a Gibbs measure concentrating on the global flat minima of the loss, i.e., the global minimizers of a Hessian-trace regularized objective. Within this framework, we further introduce layerwise normalized ZOLD (LN-ZOLD) that addresses the layerwise scale heterogeneity of pretrained LLMs at no additional memory or forward evaluation. On the theoretical side, we establish non-asymptotic Wasserstein-1 error bounds for ZOLD with respect to the flatness-biased Gibbs measure. Empirically, ZOLD and LN-ZOLD improve over MeZO in accuracy across OPT-13B/30B/66B, LLaMA3, and Phi-2, while remaining substantially faster than HiZOO and FZOO at comparable accuracy.

1 INTRODUCTION

Large language models (LLMs) have achieved strong performance across diverse natural language processing tasks, including question answering, reasoning, and text generation (Brown et al., 2020; Radford et al., 2019; Liu et al., 2019b). However, adapting pretrained LLMs to specific downstream tasks typically requires fine-tuning, which can be computationally and memory intensive. While inference can often be performed within modest memory budgets, fine-tuning is substantially more expensive. For example, first-order (FO) optimizers such as Adam (Kingma & Ba, 2015) must store, in addition to the model parameters, gradients, optimizer states, and activations required for backpropagation (Liu et al., 2024), resulting in memory costs several times larger than those of inference (e.g., over 600 GB for OPT-30B).

Zeroth-order (ZO) optimization has recently emerged as a memory-efficient alternative for LLM fine-tuning (Malladi et al., 2023). ZO methods estimate gradients from forward evaluations alone, removing the need for backpropagation and reducing memory consumption to near-inference levels. Malladi et al. (2023) introduced MeZO, which established this paradigm for LLMs and has since served as the reference point for a growing line of work. The vast majority of subsequent ZO methods for LLM fine-tuning, including (Zhao et al., 2025; Liu et al., 2025; Chen et al., 2025; Yue et al., 2025; Yu et al., 2025; Guo et al., 2025; Tan et al., 2025; Dang et al., 2026; Behric et al., 2026), share a common structure: they replace the gradient in an SGD-style update with a refined ZO estimator,

$$\theta_{k+1} = \theta_k - \lambda \widetilde{G}(\theta_k, Z_{k+1}, X_{k+1}), \tag{1}$$

where X_{k+1} is a random variable representing data, $Z_{k+1} \sim \mathcal{N}(0, I_d)$ is the Gaussian perturbation used to construct the estimator, and $\widetilde{G}$ represents a modified ZO estimator incorporating mechanisms such as sparsification, low-rank projection, curvature correction, or variance reduction. We refer to this collection of methods as the *ZO-SGD family*. The progress within this family has been driven almost entirely by improvements to the estimator $\widetilde{G}$, while the underlying optimizer remains SGD.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite this rapid progress, the entire ZO-SGD family inherits a structural limitation: by retaining the SGD form in (1), every member of the family is bound to the convergence behavior of SGD. In a general non-convex setting, SGD-style optimization typically guarantees convergence only to stationary points[1] and provides no mechanism for global exploration of the loss landscape. Since this limitation arises from the SGD framework, it cannot be addressed by refining the estimator $\widetilde{G}$. Furthermore, prior work within the ZO-SGD family has largely focused on mitigating the high variance of ZO gradient estimators, either through improved estimator designs (Gautam et al., 2024; Yue et al., 2025; Behric et al., 2026), or through adaptive scaling and second-order corrections (Dang et al., 2026; Zhao et al., 2025). While these techniques achieve their intended objectives, they often introduce additional memory cost or require extra forward evaluations. An underexplored issue is the layerwise scale heterogeneity of deep models: standard ZO estimators inherently apply a uniform perturbation scale across all parameters, which makes them ineffective at handling such heterogeneity. Existing attempts to address this (Tan et al., 2025) require an additional optimization step for computing layerwise projections, resulting in extra computational overhead. A detailed literature review and further discussion are provided in Appendix A.

Table 1: Comparison of ZO methods. **Non-convex**: whether the analysis covers non-convex objectives. **Global guarantee**: whether guarantees are established to global optima rather than to stationary points. **Flat minima**: whether the analysis explicitly accounts for flat minima.

Algorithm	Non-convex	Global guarantee	Flat minima
MeZO (Malladi et al., 2023)	✓	✓	×
ZO (Zhang et al., 2025)	×	×	✓
HiZOO (Zhao et al., 2025)	✓	×	×
FZOO (Dang et al., 2026)	✓	×	×
ZOLD (Ours)	✓	✓	✓

In this paper, we propose a novel family of zeroth-order methods for LLM fine-tuning, which we call *Zeroth-Order Langevin Dynamics* (ZOLD). Our approach can be viewed as a zeroth-order analogue of stochastic gradient Langevin dynamics (SGLD) (Welling & Teh, 2011; Raginsky et al., 2017). Concretely, the update takes the form

$$\theta_{k+1} = \theta_k - \lambda \widetilde{G}(\theta_k, Z_{k+1}, X_{k+1}) + \sqrt{2\lambda\beta^{-1}}\,\xi_{k+1}, \tag{2}$$

where $(\xi_k)_{k\in\mathbb{N}}$ is an independent sequence of standard d-dimensional Gaussian random variables, independent of the perturbation Z_{k+1}. In contrast to the ZO-SGD family in (1), the injected Langevin noise ξ_{k+1} provides a structural mechanism for global exploration, enabling error analysis with respect to global optima rather than to stationary points. Within the proposed ZOLD framework, we further introduce *Layerwise Normalized ZOLD* (LN-ZOLD), which addresses the layerwise scaling imbalance without requiring additional memory or forward evaluations. Our main contributions are summarized as follows:

- We open a new direction for zeroth-order optimization in LLM fine-tuning by departing from the ZO-SGD family and proposing ZOLD, a new framework based on SGLD that improves both theoretical analysis and empirical performance.
- Through a precise coupling between the perturbation scale η and the Langevin scale β, we establish non-asymptotic Wasserstein-1 error bounds for ZOLD (Theorem 4.4) with respect to a Gibbs measure concentrating on the global minima of a Hessian-trace-regularized objective. Our analysis goes beyond prior work that is limited to local convergence to stationary points, relies on convexity assumptions, or does not consider flat minima; see Table 1.
- Within the ZOLD framework, we further introduce LN-ZOLD, a layerwise normalized variant that yields an adaptive step size at no additional computational cost. Inspired by tamed schemes for SDEs, it improves stability under heterogeneous parameter scales in high-dimensional settings.
- We validate ZOLD and LN-ZOLD on OPT-13B/30B/66B, LLaMA3, and Phi-2 across classification, multiple-choice, and generation tasks. Our methods incur memory footprint and per-iteration forward cost comparable to MeZO, making them substantially faster in wall-clock time than strong baselines, such as HiZOO (Zhao et al., 2025) and FZOO (Dang et al.,

[1]Global convergence is achievable under strong structural assumptions on the loss such as the Polyak-Łojasiewicz condition (Karimi et al., 2016). Malladi et al. (2023) establish a global convergence result for MeZO under such an assumption.

2026), which rely on multiple perturbations or extra memory, while exceeding MeZO in accuracy and remaining competitive with these more expensive methods; see Figure 2.

2 PRELIMINARIES

Notation. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a fixed probability space. The probability law of a random variable Y is denoted by $\mathcal{L}(Y)$. Fix integers $d, m \geq 1$. The Euclidean scalar product is denoted by $\langle \cdot, \cdot \rangle$, with $\| \cdot \|$ standing for the corresponding norm. The notation $\widetilde{\mathcal{O}}(\cdot)$ is used to omit logarithmic factors in the relevant parameters. Let $u : \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable function, and we denote its gradient by ∇u. Let $\mathcal{P}(\mathbb{R}^d)$ be the set of probability measures on $\mathcal{B}(\mathbb{R}^d)$, and for $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, let $\mathcal{A}(\mu, \nu)$ denote the set of probability measures ϱ on $\mathcal{B}(\mathbb{R}^{2d})$ such that its respective marginals are μ and ν. For any μ and $\nu \in \mathcal{P}(\mathbb{R}^d)$, the Wasserstein distance of order 1 is defined as

$$W_1(\mu, \nu) = \inf_{\varrho \in \mathcal{A}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \|x - z\| \, \mathrm{d}\varrho(x, z).$$

2.1 LANGEVIN DYNAMICS FOR GLOBAL EXPLORATION

Langevin dynamics offers a stochastic mechanism for global exploration of non-convex optimization landscapes. Consider the overdamped Langevin diffusion

$$\mathrm{d}Y_t = -\nabla u(Y_t) \, \mathrm{d}t + \sqrt{2\beta^{-1}} \, \mathrm{d}B_t, \tag{3}$$

where $\beta > 0$ denotes the inverse temperature and $(B_t)_{t \geq 0}$ is a d-dimensional Brownian motion. Under mild regularity conditions, this dynamics admits the Gibbs invariant measure

$$\pi_\beta^u(\theta) \propto \exp(-\beta u(\theta)),$$

which concentrates around the global minimizers of u for sufficiently large β (Hwang, 1980). The Euler–Maruyama discretization of (3) with a stochastic gradient yields SGLD, which serves both as a sampling algorithm (Welling & Teh, 2011) and as a global optimizer (Raginsky et al., 2017). Non-asymptotic guarantees for SGLD in non-convex settings have been established under various assumptions (Raginsky et al., 2017; Xu et al., 2018; Majka et al., 2020; Chau et al., 2021; Zhang et al., 2023), and recent work has adapted SGLD to target flat minima (Bruno et al., 2026). However, existing Langevin-based methods rely on first-order gradient information, which limits their applicability in memory-constrained settings such as LLM fine-tuning.

2.2 ZEROTH-ORDER OPTIMIZATION, IMPLICIT REGULARIZATION, AND FLAT MINIMA

We consider the non-convex stochastic optimization problem

$$\min_{\theta \in \mathbb{R}^d} u(\theta) := \mathbb{E}[U(\theta, X)], \tag{4}$$

where $U : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}$ is a measurable function with $\mathbb{E}[|U(\theta, X)|] < \infty$ for all $\theta \in \mathbb{R}^d$, and X is a random variable with distribution $\mathcal{L}(X)$. Zeroth-order (ZO) optimization aims to solve (4) when ∇u is unavailable or expensive to compute, relying only on function evaluations $U(\theta, X)$ without access to gradient information. To this end, ZO methods replace gradients with finite-difference estimators. A standard choice is the two-point estimator (Spall, 1992; Nesterov & Spokoiny, 2017)

$$G_\eta(\theta, Z, X) := \frac{U(\theta + \eta Z, X) - U(\theta - \eta Z, X)}{2\eta} \, Z, \tag{5}$$

where $\eta > 0$ is a smoothing parameter and $Z \sim \mathcal{N}(0, I_d)$. This estimator is *biased* for the true gradient ∇f but is an unbiased estimator for ∇u_η:

$$\mathbb{E}_{Z,X}[G_\eta(\theta, Z, X)] = \nabla u_\eta(\theta), \tag{6}$$

$$\text{with } u_\eta(\theta) = \mathbb{E}[u(\theta + \eta Z)]. \tag{7}$$

Most existing analyses of ZO methods for LLM fine-tuning (Malladi et al., 2023; Zhao et al., 2025; Dang et al., 2026; Tan et al., 2025; Behric et al., 2026) are conducted with respect to the original objective u, implicitly treating the zeroth-order estimator as an approximation of ∇u in the small-η

regime. This perspective neglects the regularization effect that finite η induces on u_η. Specifically, for sufficiently smooth u, Taylor's theorem yields

$$u_\eta(\theta) = u(\theta) + \frac{\eta^2}{2}\operatorname{tr}(H(\theta)) + \mathcal{O}(\eta^4), \tag{8}$$

where $\operatorname{tr}(H(\theta))$ denotes the trace of the Hessian of u. Gaussian smoothing thus introduces an implicit curvature-dependent regularization term controlled by η, biasing optimization dynamics toward regions of low average curvature. This connects to the flat minima hypothesis (Hochreiter & Schmidhuber, 1997), which associates such regions with improved generalization (Neyshabur et al., 2017; Damian et al., 2021). Motivated by this, we study the Hessian-trace regularized objective

$$v(\theta) = u(\theta) + \frac{\eta^2}{2}\operatorname{tr}(H(\theta)), \tag{9}$$

whose minimizers represent a trade-off between low loss and low curvature and are referred to as *global flat minima*, i.e., $\arg\min_{\theta\in\mathbb{R}^d} v(\theta)$. The associated flatness-biased Gibbs measure

$$\pi^\star_{\beta,\eta}(\theta) \propto \exp(-\beta v(\theta)) \tag{10}$$

concentrates around these global flat minima for sufficiently large β, providing the target distribution for our analysis. Recent work (Zhang et al., 2025) provides theoretical evidence that ZO methods can favor flat minima under convexity and third-order smoothness assumptions on u. A detailed derivation of the relationship between u_η and $v(\theta)$, including the approximation error, is provided in Appendix C.

3 The ZOLD Framework

In this section, we introduce the ZOLD framework, its basic form *vanilla ZOLD*, and a variant *layerwise normalized ZOLD* (LN-ZOLD) along with its properties.

Formally, let $\theta_0 \in \mathbb{R}^d$ be the initialization, and let $(X_k)_{k\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathcal{L}(X)$, $(Z_k)_{k\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$, and $(\xi_k)_{k\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d)$ be mutually independent sequences. We recall the ZOLD framework in (2): at iteration k,

$$\theta_{k+1} = \theta_k - \lambda\,\widetilde{G}(\theta_k, Z_{k+1}, X_{k+1}) + \sqrt{2\lambda\beta^{-1}}\,\xi_{k+1}, \tag{11}$$

where $\lambda > 0$ is the step size, $\beta > 0$ is the inverse temperature, $\widetilde{G}$ is a zeroth-order gradient estimator, Z_{k+1} is the perturbation noise used to construct $\widetilde{G}$, and ξ_{k+1} is the injected Langevin noise.

3.1 Vanilla ZOLD

The most common choice of $\widetilde{G}$ is the standard two-point Gaussian-smoothed estimator (5) with smoothing radius $\eta > 0$. Substituting $\widetilde{G} = G_\eta$ in (11) yields the *vanilla ZOLD* update

$$\theta_{k+1} = \theta_k - \lambda\,G_\eta(\theta_k, Z_{k+1}, X_{k+1}) + \sqrt{2\lambda\beta^{-1}}\,\xi_{k+1}, \tag{12}$$

with the coupling

$$\eta = \beta^{-\frac{1+\tau}{4}}, \qquad \text{for a fixed } \tau \in (0,1). \tag{13}$$

Vanilla ZOLD features a structured interaction between two sources of randomness: the Gaussian perturbation Z_k used in the zeroth-order estimator G_η (scaled by η) and the Langevin noise ξ_k (scaled by β^{-1}). Intuitively, the vanilla ZOLD update (12) can be interpreted as a zeroth-order discretization of the overdamped Langevin dynamics (3), where the true gradient ∇u is replaced by ∇u_η due to (6). Therefore, the invariant measure of the vanilla ZOLD dynamics is $\pi_\beta(\mathrm{d}\theta) \propto \exp(-\beta u_\eta(\theta))$.

Our goal is to control the discrepancy between π_β and the target flatness-biased Gibbs measure $\pi^\star_{\beta,\eta}$ in (10). To this end, we write

$$\pi_\beta(\mathrm{d}\theta) = \frac{Z^\star_{\beta,\eta}}{Z_\beta}\exp\big(-\beta(u_\eta(\theta) - v(\theta))\big)\,\pi^\star_{\beta,\eta}(\mathrm{d}\theta), \tag{14}$$

where $Z_\beta, Z^\star_{\beta,\eta}$ are the corresponding normalization constants. By Lemma D.11, $|u_\eta - v| \lesssim \mathcal{O}(\eta^4 d^2)$, so the exponential factor in (14) is governed by $\beta\eta^4 d^2$. The coupling (13) ensures this factor remains controllable: substituting $\eta = \beta^{-(1+\tau)/4}$ gives $\beta\eta^4 = \beta^{-\tau}$, which can be made arbitrarily small by choosing β sufficiently large.

Remark 3.1. *A natural concern is whether increasing β, required to align π_β with $\pi^\star_{\beta,\eta}$, also suppresses the flatness regularization $\frac{\eta^2}{2}\operatorname{tr}(H(\theta))$ in v. This is not the case. The regularization term decays at order $\eta^2 = \beta^{-(1+\tau)/2}$,* strictly slower *than the remainder $|u_\eta - v|$ of order $\eta^4 = \beta^{-(1+\tau)}$ in (14). The regularization thus remains significant relative to the discrepancy controlled by (14), so the bias toward flat minima and the closeness of π_β to $\pi^\star_{\beta,\eta}$ coexist under the coupling.*

From a practical perspective, this coupling reduces hyperparameter tuning to a single parameter β, since the smoothing parameter η is automatically determined and does not require separate tuning or scheduling. The parameter $\tau \in (0, 1)$ in (13) is kept fixed throughout (set to $\tau = 0.5$ in all experiments) and plays a technical role in the theoretical analysis.

3.2 Layerwise Normalized ZOLD (LN-ZOLD)

We introduce *layerwise normalized ZOLD* (LN-ZOLD), a variant of ZOLD that addresses two key challenges in zeroth-order LLM fine-tuning: layerwise scale heterogeneity in deep models and outlier parameter magnitudes, such as super weights in pretrained LLMs (Zhou et al., 2025; Yu et al., 2024).

We decompose $\theta = (\theta^{(1)}, \dots, \theta^{(L)})$ into L layerwise blocks with $\theta^{(l)} \in \mathbb{R}^{d_l}$ and $\sum_l d_l = d$, and write $G^{(l)}_\eta(\theta, Z, X) \in \mathbb{R}^{d_l}$ for the corresponding sub-vector of $G_\eta(\theta, Z, X)$. Note that G_η itself remains unchanged: the perturbation Z is applied jointly across all parameters, preserving the unbiased property $\mathbb{E}[G_\eta] = \nabla u_\eta$ in (6). Then, the *layerwise normalized estimator* $G_{\eta,\lambda}$ is obtained by normalizing each layerwise component:

$$G^{(l)}_{\eta,\lambda}(\theta, Z, X) = \frac{G^{(l)}_\eta(\theta, Z, X)}{1 + \sqrt{\lambda}\,\|\theta^{(l)}\|}, \qquad l = 1, \dots, L. \tag{15}$$

Substituting $\widetilde{G} = G_{\eta,\lambda}$ in (11) yields the ***LN-ZOLD*** update

$$\theta_{k+1} = \theta_k - \lambda\, G_{\eta,\lambda}(\theta_k, Z_{k+1}, X_{k+1}) + \sqrt{2\lambda\beta^{-1}}\,\xi_{k+1}, \tag{16}$$

under the same coupling (13).

LN-ZOLD has three desirable properties. First, its design is inspired by *tamed schemes* (Hutzenthaler et al., 2012; Sabanis, 2013), well-established numerical methods for stable SDE approximation. This perspective leads to more stable training dynamics in practice. In contrast to prior ZO approaches that primarily emphasize variance reduction and estimator efficiency (Gautam et al., 2024; Yue et al., 2025; Behric et al., 2026), LN-ZOLD explicitly incorporates stability mechanisms from numerical analysis for SDEs, mitigating training instabilities associated with large or heterogeneous updates. Second, the layerwise normalization induces a layerwise adaptive step size: the effective step size for layer l scales as $\lambda/(1 + \sqrt{\lambda}\,\|\theta^{(l)}\|)$, which is achieved without additional forward evaluations or memory beyond those of vanilla ZOLD. Third, for layers with large parameter norms $\|\theta^{(l)}\|$ such as those containing super weights in pretrained LLMs (Zhou et al., 2025; Yu et al., 2024), the normalization factor $1 + \sqrt{\lambda}\|\theta^{(l)}\|$ induces a scale-dependent damping of the update magnitude. While in typical regimes $\sqrt{\lambda}\|\theta^{(l)}\|$ remains smaller than one, this mechanism still reduces the relative update size for large-norm layers, contributing to improved stability during fine-tuning. In regimes where $\sqrt{\lambda}\|\theta^{(l)}\|$ becomes non-negligible, this effect becomes more pronounced and approaches an inverse scaling with the parameter norm.

4 Theoretical Analysis and Practical Strategy

In this section, we establish non-asymptotic W_1 error bounds between the law of the ZOLD iterates and the target Gibbs measure $\pi^\star_{\beta,\eta} \propto \exp(-\beta v(\theta))$, together with the corresponding iteration complexities. We also introduce a practical strategy based on β scheduling.

4.1 ASSUMPTIONS

We first impose independence of the random sources with unbiasedness of the stochastic gradient.

Assumption 4.1 (Independence and unbiasedness). *The random variables θ_0, $(Z_k)_{k\in\mathbb{N}}$, $(X_k)_{k\in\mathbb{N}}$, and $(\xi_k)_{k\in\mathbb{N}}$ in (11) are mutually independent, $\mathbb{E}[\|\theta_0\|^4] < \infty$, and $\mathbb{E}[\|X\|^4] < \infty$. The stochastic gradient is unbiased: $\mathbb{E}[\nabla_\theta U(\theta, X)] = \nabla u(\theta)$ and $\mathbb{E}[U(\theta, X)] = u(\theta)$ for all $\theta \in \mathbb{R}^d$.*

Next, we impose third-order smoothness of u together with Lipschitz continuity of the stochastic gradient.

Assumption 4.2 (Regularity). *The function $u : \mathbb{R}^d \to \mathbb{R}$ is three-times continuously differentiable and third-order smooth. Moreover, there exist $L_1, L_2 > 0$ such that, for all $x, x' \in \mathbb{R}^m$ and $\theta, \theta' \in \mathbb{R}^d$,*

$$\|\nabla_\theta U(\theta, x) - \nabla_{\theta'} U(\theta', x)\| \le L_1 \|\theta - \theta'\|,$$
$$\|\nabla_\theta U(\theta, x) - \nabla_\theta U(\theta, x')\| \le L_2(1 + \|\theta\|)\|x - x'\|.$$

The third-order smoothness is needed to characterize the Hessian-trace regularization in (9) and is standard in studies of implicit regularization (Arora et al., 2022; Wen et al., 2023), sharp minima escape (Ahn et al., 2024), and flat-minima analyses for zeroth- and first-order methods (Zhang et al., 2025; Bruno et al., 2026). Finally, we impose a dissipativity condition that ensures stability of the Langevin dynamics, a commonly used assumption in non-asymptotic analyses of Langevin-based algorithms (Raginsky et al., 2017; Xu et al., 2018; Deng et al., 2022; Futami & Fujisawa, 2023).

Assumption 4.3 (Dissipativity). *There exist $a > 0$ and $b \ge 0$ such that, for all $\theta \in \mathbb{R}^d$ and $x \in \mathbb{R}^m$,*

$$\langle \nabla_\theta U(\theta, x), \theta \rangle \ge a\|\theta\|^2 - b.$$

4.2 NON-ASYMPTOTIC W_1 ERROR BOUNDS

Our main result establishes non-asymptotic W_1 error bounds between the law of the vanilla ZOLD iterates and the target Gibbs measure $\pi^\star_{\beta,\eta}$, which concentrates on global flat minima.

Theorem 4.4. *Let Assumptions 4.1, 4.2, and 4.3 hold. For any $\epsilon > 0$, let $\beta > 0$ be sufficiently large, and choose*

$$\lambda \le \mathcal{O}(\epsilon^2 \dot{c}^2 \beta^{(1+\tau)/2}/d^3) \wedge \lambda_{\max}, \quad \eta = \beta^{-(1+\tau)/4},$$

for $\tau \in (0, 1)$, with $\lambda_{\max}$ given in (36), and $\dot{c}$ is the contraction constant of the Langevin diffusion (3), given in Eberle et al. (2019, Theorem 2.2). Then $W_1(\mathcal{L}(\theta_k), \pi^\star_{\beta,\eta}) \le \epsilon$ after

$$k \ge \widetilde{\mathcal{O}}\left(\frac{d^3}{\epsilon^2 \beta^{(1+\tau)/2}} \mathrm{poly}(1/\dot{c})\right) \text{ iterations.} \tag{17}$$

To the best of our knowledge, Theorem 4.4 provides the first non-asymptotic guarantee with respect to *global flat minima* for zeroth-order methods in non-convex optimization under dissipativity and smoothness assumptions. Prior non-asymptotic analyses for zeroth-order methods either provide global guarantees but do not account for flat minima (Malladi et al., 2023), account for flat minima only under convexity (Zhang et al., 2025), or remain local in the form of guarantees to stationary points (Zhao et al., 2025; Dang et al., 2026); see Table 1. An overview of the proof strategy for Theorem 4.4 is given at the beginning of Appendix D, while the full proof is deferred to Appendix D.3.

4.3 HYPERPARAMETER TUNING GUIDE VIA β ANNEALING

In practice, inverse-temperature scheduling is widely used in SGLD and its variants for optimization purposes (Chen et al., 2016; Deng et al., 2020a; 2019). The principle is to start with a small β to encourage global exploration of the loss landscape, and to gradually increase β to drive the dynamics toward exploitation, thereby concentrating around the global minima. Among various possible schedules, we adopt the simple geometric form

$$\beta_k = \beta_0^{1-s_k} \beta_K^{s_k}, \qquad s_k := k/K, \qquad k = 0, \dots, K, \tag{18}$$

where K is the total number of training iterations. This schedule is log-linear and gradually increases β_k from β_0 to β_K, thereby reducing the scale of the injected Langevin noise. Through the

coupling (13), the smoothing radius then becomes time-varying, $\eta_k = \beta_k^{-(1+\tau)/4}$, so that β_k jointly controls the global exploration and the strength of the curvature regularization. The schedule thus extends the coupling (13) to a dynamic setting while preserving the single-parameter structure of ZOLD. We note that annealing in ZOLD is adopted as a practical implementation choice: our theoretical analysis is restricted to the fixed-temperature regime, and extending the results to annealed variants is left for future work.

5 Experiments

5.1 An Illustrative Two-Dimensional Example

We first compare the optimization behavior of GD, MeZO, and ZOLD on a two-dimensional objective with one narrow sharp minimum, one local flat minimum, and one lower global flat minimum. The exact landscape and implementation details are provided in Appendix E.2. Figure 1 shows the trajectories from the same initialization, with identical learning rate and iteration budget. GD follows the local gradient and is attracted to the sharp basin. Through zeroth-order smoothing, MeZO inherits a flatness bias and avoids the sharp basin, but settles in the local flat minimum due to its lack of global exploration. ZOLD has both the flatness bias and global exploration. As a result, it escapes the local basin and reaches the global flat minimum. We also provide sharpness measures at the converged iterates, including the Hessian trace, in Figure 4 of Appendix E.2.

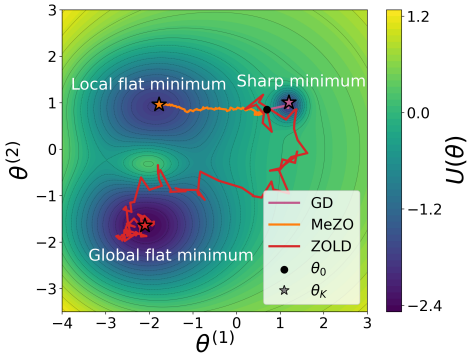


Figure 1: Trajectories on a two-dimensional objective. Starting from the same initialization θ_0, GD, MeZO, and ZOLD follow different optimization trajectories and terminate at distinct basins, with θ_K indicated by stars.

5.2 Experimental Setup

Datasets. We evaluate on three groups of tasks. The classification tasks include SST-2 (Socher et al., 2013), RTE (Dagan et al., 2005; Bentivogli et al., 2009), CB (De Marneffe et al., 2019), BoolQ (Clark et al., 2019), WSC (Levesque et al., 2012), and WiC (Pilehvar & Camacho-Collados, 2019). The multiple-choice tasks are COPA (Roemmele et al., 2011) and ReCoRD (Zhang et al., 2018). The generation tasks are SQuAD (Rajpurkar et al., 2016) and DROP (Dua et al., 2019).

Models and Baselines. We evaluate ZOLD and LN-ZOLD against MeZO (Malladi et al., 2023), FZOO (Dang et al., 2026), and the HiZOO family (Zhao et al., 2025). For autoregressive experiments on OPT (Zhang et al., 2022), LLaMA3 (Grattafiori et al., 2024), and Phi-2 (Javaheripi et al., 2023), we compare with MeZO, HiZOO-L, and FZOO. Here, HiZOO-L is the low-rank version of HiZOO, which reduces Hessian-memory overhead by factorizing the Hessian inverse approximation into low-rank matrices.

Training Details. To ensure comparability with prior zeroth-order fine-tuning results, we adopt the task setup and evaluation protocol of FZOO (Dang et al., 2026). We evaluate ZOLD and LN-ZOLD under the same protocol, and use the baseline results reported in FZOO (Dang et al., 2026) for MeZO, FZOO, and the HiZOO variants. Following this protocol, we report the best performance across the considered hyperparameter settings. Additional experimental details and hyperparameter configurations are provided in Appendix E.1.

5.3 Main Results

We evaluate ZOLD and LN-ZOLD on autoregressive language models across classification, multiple-choice, and generation tasks. Table 2 reports the main results on OPT-13B, LLaMA3, and Phi-2. ZOLD and LN-ZOLD achieve the two highest average scores on each of the three model families, improving over MeZO by +2.1, +2.8, and +4.5 points on OPT-13B, LLaMA3, and Phi-2, respectively. The improvements are observed across classification, multiple-choice, and generation

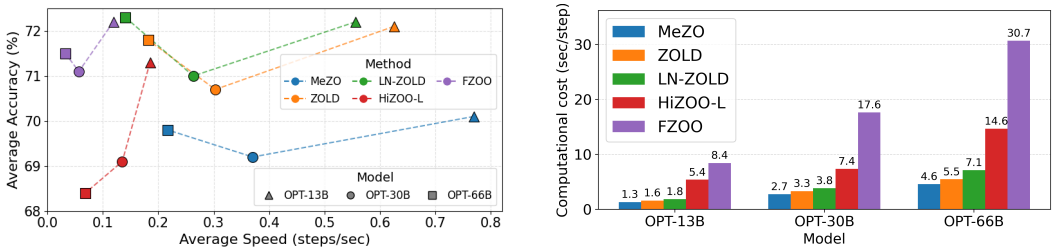
Table 2: Performance comparison on autoregressive language models. We report results on classification (SST-2, RTE, CB, BoolQ, WSC, WiC), multiple-choice (COPA, ReCoRD), and generation (SQuAD, DROP) tasks. The best result is highlighted in **bold** and the second-best is underlined.

Model	Method	SST-2	RTE	CB	BoolQ	WSC	WiC	COPA	ReCoRD	SQuAD	DROP	Avg.
OPT-13B	MeZO	91.4	66.1	66.0	67.6	63.5	59.4	88.0	**81.7**	84.7	**30.9**	69.9
	HiZOO-L	92.1	68.2	67.9	66.6	**65.4**	59.4	**89.0**	81.1	63.6	22.7	67.6
	FZOO	**93.7**	71.1	69.6	72.2	63.5	**60.5**	87.0	81.0	**84.8**	28.7	71.2
	ZOLD	93.0	**71.5**	69.6	72.3	64.4	59.4	**89.0**	81.6	83.9	30.8	71.6
	LN-ZOLD	92.9	70.0	**73.2**	**72.8**	**65.4**	**60.5**	**89.0**	81.4	83.6	**30.9**	**72.0**
LLaMA3	MeZO	92.2	74.4	69.6	76.7	63.5	57.8	88.0	85.6	86.7	57.1	75.2
	HiZOO-L	**94.3**	75.1	69.6	77.1	63.5	57.7	89.0	85.6	87.7	49.4	74.9
	FZOO	**94.3**	**77.6**	69.6	81.8	**65.4**	60.8	88.0	85.3	**87.9**	56.5	76.7
	ZOLD	93.7	77.3	**71.4**	81.8	63.5	**65.0**	**92.0**	**85.8**	87.1	62.4	**78.0**
	LN-ZOLD	93.1	76.2	**71.4**	**82.3**	62.5	**65.0**	**92.0**	85.7	87.3	**62.6**	77.8
Phi-2	MeZO	86.6	67.1	75.0	72.4	59.6	54.4	86.0	71.7	85.7	37.8	69.6
	HiZOO-L	88.9	68.9	75.2	72.0	**62.4**	59.2	86.0	72.1	85.7	36.2	70.7
	FZOO	87.4	70.4	83.9	79.3	61.5	56.7	86.0	72.0	86.7	37.4	72.1
	ZOLD	**90.4**	**75.5**	**85.7**	**80.1**	58.7	60.7	**90.0**	72.6	87.0	40.2	**74.1**
	LN-ZOLD	89.6	71.1	82.1	79.5	59.6	**62.5**	**90.0**	**72.7**	**87.7**	**40.7**	73.6

tasks rather than being concentrated in a single task category. We further report results on OPT-125M/2.7B/13B over three random seeds in Appendix F, including comparisons with first-order full fine-tuning (FO) and FO-LoRA, where ZOLD and LN-ZOLD remain competitive with ZO baselines.

5.4 Accuracy and Efficiency at Scale

We next examine whether the advantages of ZOLD and LN-ZOLD become more pronounced at larger model scales while preserving the computational efficiency of zeroth-order fine-tuning. Table 3 examines scalability on larger OPT models. LN-ZOLD ranks second on OPT-30B and first on OPT-66B, with ZOLD ranking second on OPT-66B. The relative improvement of LN-ZOLD over MeZO is larger at OPT-66B than at OPT-30B, while preserving the computational efficiency of ZO fine-tuning.



(a) Speed–accuracy trade-off

(b) Per-step computational cost

Figure 2: Average speed–accuracy trade-off and per-step computational cost across evaluated tasks using OPT-13B, OPT-30B, and OPT-66B.

Figure 2 compares the speed–accuracy trade-off (Figure 2a) and per-step computational cost (Figure 2b) of zeroth-order methods on SST-2, RTE, WSC, and WiC across OPT-13B, OPT-30B, and OPT-66B. Speed is measured as steps per second from the wall-clock time of 100 optimization steps, and computational cost is reported as the corresponding seconds per step.

In Figure 2a, ZOLD (orange) and LN-ZOLD (green) consistently cluster in the upper-right region of the speed–accuracy plane, indicating that they are simultaneously fast and accurate, whereas MeZO (blue) is fast but less accurate and FZOO (purple) and HiZOO-L (red) reach competitive accuracy at a speed several times slower. ZOLD improves over MeZO in average accuracy across all three model scales while incurring only a small per-step overhead (Figure 2b: 1.3 $\rightarrow$ 1.6 sec/step on OPT-13B, 4.6 $\rightarrow$ 5.5 sec/step on OPT-66B). Compared with HiZOO-L and FZOO, ZOLD attains comparable accuracy at a substantially lower per-step cost: on OPT-66B, ZOLD reaches 71.8% accuracy at 0.20

Table 3: Performance comparison on larger OPT models across classification tasks. The best result is highlighted in **bold** and the second-best is underlined.

	OPT-30B					OPT-66B				
Method	**SST-2**	**RTE**	**WSC**	**WiC**	**Avg.**	**SST-2**	**RTE**	**WSC**	**WiC**	**Avg.**
MeZO	90.6	66.4	**63.5**	56.3	69.2	91.2	65.7	63.5	58.9	69.8
HiZOO-L	90.3	69.3	**63.5**	53.4	69.1	88.9	66.4	59.6	58.6	68.4
FZOO	91.2	69.3	**63.5**	**60.2**	**71.1**	93.6	**74.4**	58.6	**59.2**	71.5
ZOLD	93.5	**70.0**	62.5	56.9	70.7	93.3	73.6	63.5	56.9	71.8
LN-ZOLD	**93.7**	69.7	**63.5**	57.1	71.0	**93.7**	73.6	**65.4**	56.6	**72.3**

Table 4: Effect of β scheduling on SST-2 using OPT-1.3B. The best result is highlighted in **bold**.

	w/o β scheduling			β scheduling
Method	10^8	10^{10}	10^{12}	$\beta_0 = 10^8,\ \beta_K = 10^{12}$
ZOLD	91.5	91.2	90.0	**91.9**
LN-ZOLD	90.7	90.7	90.8	**92.0**

steps/sec, against FZOO's 71.5% at 0.04 steps/sec. LN-ZOLD provides further accuracy gains on larger models, attaining the best accuracy on OPT-66B while remaining substantially faster than HiZOO-L and FZOO. Beyond per-step cost, Appendix F shows that ZOLD and LN-ZOLD reach their best-accuracy checkpoints on OPT-125M/2.7B/13B 2.3–8.6× faster than FZOO with 2.4–6.1× fewer total FLOPs, at comparable accuracy and MeZO-level peak GPU memory.

5.5 Effect of the β–η Coupling and β Annealing

We evaluate the effects of the β–η coupling and β annealing on SST-2 using OPT-1.3B. Figure 3 shows the performance of ZOLD for fixed $\beta = 10^{10}$ across different values of τ, which controls the coupling (13) between β and η. ZOLD maintains high accuracy for $\tau \in (0, 1)$, while performance degrades outside this range, empirically supporting the coupling regime used in our analysis. We then compare the annealing schedule (18) with fixed-β configurations in Table 4. The annealed variants achieve better accuracy for both ZOLD and LN-ZOLD, indicating that keeping β fixed throughout training is suboptimal. Increasing β over training shifts the dynamics from exploration under stronger Langevin noise to exploitation under reduced noise.

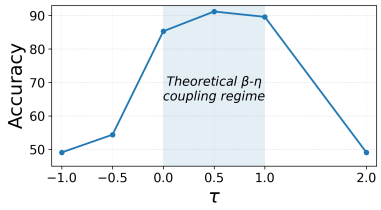


Figure 3: Sensitivity of ZOLD to the coupling parameter τ.

6 Conclusion

We introduced ZOLD, a new family of zeroth-order methods for LLM fine-tuning that departs from the ZO-SGD family by combining zeroth-order gradient estimation with Langevin dynamics. The injected Langevin noise provides a stochastic mechanism for global exploration, enabling error analysis with respect to global optima rather than to stationary points. Within the framework, we further proposed *layerwise normalized ZOLD* (LN-ZOLD), which incorporates a layerwise normalization to address the layerwise scale heterogeneity and outlier-magnitude parameters of pretrained LLMs without additional forward evaluations or memory. On the theoretical side, we established non-asymptotic W_1 error bounds for ZOLD with respect to a Gibbs measure concentrating on the global minima of a Hessian-trace regularized objective. Empirically, ZOLD and LN-ZOLD achieve a favorable speed–accuracy trade-off across OPT-13B/30B/66B, LLaMA3, and Phi-2, improving over MeZO in accuracy while remaining faster than HiZOO and FZOO.

Limitations and Future Directions. While our results establish a principled foundation, several limitations remain. First, our theoretical analysis focuses on the constant-parameter regime, where the inverse temperature and perturbation scale are fixed and appropriately coupled. In practice, annealing strategies, where these parameters evolve over time, are often used to further improve exploration and concentration properties. Providing a rigorous analysis of such annealed schemes for zeroth-order Langevin dynamics is beyond the scope of this work and would require a dedicated theoretical treatment. Second, LN-ZOLD is one specific variant of the ZOLD framework rather than an optimal design for LLM fine-tuning. The fact that LN-ZOLD does not uniformly outperform vanilla ZOLD across our experiments suggests that other structural properties of LLM fine-tuning, not captured by layerwise normalization, may also play a role. Designing such variants of ZOLD is a natural direction for future work.

AI USE STATEMENT

In this work, we have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate or prove mathematical claims, propose or refine hypotheses, design or provide feedback on the research methodology or experiments, implement the proposed methods, generate, clean, or reformat datasets, or interpret results. The remaining tasks with required disclosure (e.g., translation and qualitative data analysis) are not applicable to this work. Additionally, we used generative AI tools to assist with creating and editing software code and with drafting and editing parts of the manuscript for clarity and readability. We have reviewed all AI-assisted work, and all AI-assisted code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI. Below, we list the tasks for which disclosure is required or recommended under the ICLR 2027 AI policy.

Required: Generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results.

Recommended: Formulate questions for surveys or interviews, create or modify scientific figures or images, suggest experimental parameters, create or edit software code, creation of artifacts, draft parts of a research paper, transcribe recordings of research material, summarize or analyse existing literature, discover research topics or identify gaps, brainstorming, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, format references, suggest a structure for a research paper, propose a title or keywords for a research paper.

REPRODUCIBILITY STATEMENT

Our implementation is built on the official FZOO codebase (Dang et al., 2026) and follows the same Python environment, data-loading pipeline, training loop, and evaluation procedure, with only ZOLD and LN-ZOLD added as new optimization methods. All additional experiments are conducted using the same codebase and evaluation protocol. For the large-model comparisons, the reported results for MeZO (Malladi et al., 2023), HiZOO-L (Zhao et al., 2025), and FZOO (Dang et al., 2026) follow the published FZOO evaluation under the same task setup, data splits, and evaluation protocol, rather than being independently rerun. We provide our source code and training/evaluation scripts as supplementary material, together with the hardware configuration, package versions, hyperparameter settings, random seeds, and execution commands required for replication.

REFERENCES

Kwangjun Ahn, Ali Jadbabaie, and Suvrit Sra. How to escape sharp minima with random perturbations. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding gradient descent on the edge of stability in deep learning. In *International Conference on Machine Learning*, pp. 948–1024. PMLR, 2022.

Lejs Deen Behric, Liang Zhang, Bingcong Li, and Kiran Koshy Thekumparampil. ConMeZO: Adaptive descent-direction sampling for gradient-free finetuning of large language models. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth PASCAL recognizing textual entailment challenge. *TAC*, 7(8):1, 2009.

François Bolley and Cédric Villani. Weighted Csiszár-Kullback-Pinsker inequalities and applications to transportation inequalities. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 14, pp. 331–352, 2005.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Stefano Bruno, Youngsik Hwang, JaeHyeon An, Sotirios Sabanis, and Dongyoung Lim. Flatness-aware stochastic gradient Langevin dynamics. In *Forty-third International Conference on Machine Learning*, 2026.

Zhijie Cai, Haolong Chen, and Guangxu Zhu. AdameZO: Adam-style zeroth-order optimizer for LLM fine-tuning without maintaining the moments. In *Forty-third International Conference on Machine Learning*, 2026.

Ngoc Huy Chau, Éric Moulines, Miklos Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient Langevin dynamics with dependent data streams: The fully nonconvex case. *SIAM Journal on Mathematics of Data Science*, 3(3):959–986, 2021.

Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-SGD: Biasing gradient descent into wide valleys. *Journal of Statistical Mechanics: Theory and Experiment*, 2019 (12):124018, 2019.

Changyou Chen, David Carlson, Zhe Gan, Chunyuan Li, and Lawrence Carin. Bridging the gap between stochastic gradient MCMC and stochastic optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pp. 1051–1060. PMLR, 2016.

Yiming Chen, Yuan Zhang, Liyuan Cao, Kun Yuan, and Zaiwen Wen. Enhancing zeroth-order fine-tuning for language models with low-rank structures. In *The Thirteenth International Conference on Learning Representations*, 2025.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pp. 2924–2936, 2019.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine learning challenges workshop*, pp. 177–190. Springer, 2005.

Alex Damian, Tengyu Ma, and Jason D Lee. Label noise SGD provably prefers flat global minimizers. *Advances in Neural Information Processing Systems*, 34:27449–27461, 2021.

Sizhe Dang, Yangyang Guo, Yanjun Zhao, Haishan Ye, Xiaodong Zheng, Guang Dai, and Ivor Tsang. FZOO: Fast zeroth-order optimizer for fine-tuning large language models towards Adam-scale speed. In *The Fourteenth International Conference on Learning Representations*, 2026.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. The CommitmentBank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pp. 107–124, 2019.

Wei Deng, Xiao Zhang, Faming Liang, and Guang Lin. An adaptive empirical Bayesian method for sparse deep learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Wei Deng, Qi Feng, Liyao Gao, Faming Liang, and Guang Lin. Non-convex learning via replica exchange stochastic gradient MCMC. In *International Conference on Machine Learning*, pp. 2474–2483. PMLR, 2020a.

Wei Deng, Guang Lin, and Faming Liang. A contour stochastic gradient Langevin dynamics algorithm for simulations of multi-modal distributions. *Advances in neural information processing systems*, 33:15725–15736, 2020b.

Wei Deng, Siqi Liang, Botao Hao, Guang Lin, and Faming Liang. Interacting contour stochastic gradient Langevin dynamics. In *International Conference on Learning Representations*, 2022.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2368–2378, 2019.

John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

Andreas Eberle, Arnaud Guillin, and Raphael Zimmer. Quantitative Harris-type theorems for diffusions and McKean–Vlasov processes. *Transactions of the American Mathematical Society*, 371 (10):7135–7173, 2019.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Futoshi Futami and Masahiro Fujisawa. Time-independent information-theoretic generalization bounds for SGLD. *Advances in Neural Information Processing Systems*, 36:8173–8185, 2023.

Xuefeng Gao, Mert Gürbüzbalaban, and Lingjiong Zhu. Global convergence of stochastic gradient Hamiltonian Monte Carlo for nonconvex stochastic optimization: Nonasymptotic performance bounds and momentum-based acceleration. *Operations Research*, 70(5):2931–2947, 2022.

Tanmay Gautam, Youngsuk Park, Hao Zhou, Parameswaran Raman, and Wooseok Ha. Variance-reduced zeroth-order methods for fine-tuning language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24, 2024.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM journal on optimization*, 23(4):2341–2368, 2013.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Wentao Guo, Jikai Long, Yimeng Zeng, Zirui Liu, Xinyu Yang, Yide Ran, Jacob R. Gardner, Osbert Bastani, Christopher De Sa, Xiaodong Yu, Beidi Chen, and Zhaozhuo Xu. Zeroth-order fine-tuning of LLMs with transferable static sparsity. In *The Thirteenth International Conference on Learning Representations*, 2025.

Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural computation*, 9(1):1–42, 1997.

Martin Hutzenthaler, Arnulf Jentzen, and Peter E. Kloeden. Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients. *The Annals of Applied Probability*, 22(4):1611–1641, 2012.

Chii-Ruey Hwang. Laplace's method revisited: Weak convergence of probability measures. *The Annals of Probability*, pp. 1177–1182, 1980.

Kevin G Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. *Advances in neural information processing systems*, 25, 2012.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3, 2023.

Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International conference on machine learning*, pp. 3100–3109. PMLR, 2019.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 795–811. Springer, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. ASAM: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International conference on machine learning*, pp. 5905–5914. PMLR, 2021.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. *KR*, 2012(13th):3, 2012.

Wei Lin, Yining Jiang, Qingyu Song, Qiao Xiang, and Hong Xu. AGZO: Activation-guided zeroth-order optimization for LLM fine-tuning. In *Forty-third International Conference on Machine Learning*, 2026.

Lewis Liu and Zhaoran Wang. One-point gradient estimators for zeroth-order stochastic gradient Langevin dynamics, 2020. OPT2020: 12th Annual Workshop on Optimization for Machine Learning Poster #96. https://opt-ml.org/oldopt/papers/2020/paper_96.pdf.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. *Advances in neural information processing systems*, 31, 2018.

Sijia Liu, Pin-Yu Chen, Xiangyi Chen, and Mingyi Hong. signSGD via zeroth-order oracle. In *International conference on learning representations*, 2019a.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019b.

Yong Liu, Zirui Zhu, Chaoyu Gong, Minhao Cheng, Cho-Jui Hsieh, and Yang You. Sparse MeZO: Less parameters for better performance in zeroth-order LLM fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zirui Liu, Guanchu Wang, Shaochen Henry Zhong, Zhaozhuo Xu, Daochen Zha, Ruixiang Ryan Tang, Zhimeng Stephen Jiang, Kaixiong Zhou, Vipin Chaudhary, Shuai Xu, et al. Winner-take-all column row sampling for memory efficient adaptation of language model. *Advances in Neural Information Processing Systems*, 36, 2024.

Mateusz B. Majka, Aleksandar Mijatović, and Łukasz Szpruch. Nonasymptotic bounds for sampling algorithms without log-concavity. *Annals of Applied Probability*, 30(4):1534–1581, 2020.

Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 36:53038–53075, 2023.

Shogo Nakakita. Non-asymptotic analysis of Langevin-type Monte Carlo algorithms. *Advances in Applied Probability*, pp. 1–38, 2025.

Emanuele Naldi, Marco Rando, Lorenzo Rosasco, and Silvia Villa. Langevin for nonconvex optimization: exact, inexact and zeroth-order. *arXiv preprint arXiv:2607.22353*, 2026.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17:527–566, 2017.

Ariel Neufeld, Ying Zhang, et al. Non-asymptotic convergence bounds for modified tamed unadjusted Langevin algorithm in non-convex setting. *Journal of Mathematical Analysis and Applications*, 543(1), 2025.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017.

Mohammad Taher Pilehvar and Jose Camacho-Collados. WiC: The word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1267–1273, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: A nonasymptotic analysis. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pp. 1674–1703. PMLR, 2017.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pp. 90–95, 2011.

Abhishek Roy, Lingqing Shen, Krishnakumar Balasubramanian, and Saeed Ghadimi. Stochastic zeroth-order discretizations of Langevin diffusions for Bayesian inference. *Bernoulli*, 28(3): 1810–1834, 2022.

Sotirios Sabanis. A note on tamed Euler approximations. *Electronic Communications in Probability*, 18:1 – 10, 2013.

M. Berk Sahin, Behzad Sharif, and Abolfazl Hashemi. Zeroth-order non-log-concave sampling with variance reduction and applications to inverse problems. In *Forty-third International Conference on Machine Learning*, 2026.

Hyunseok Seung, Jaewoo Lee, and Hyunsuk Ko. Low-rank curvature for zeroth-order optimization in LLM fine-tuning. In *AAAI*, pp. 25235–25242, 2026.

Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.

James C Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE transactions on automatic control*, 37(3):332–341, 1992.

Liyan Tan, Yequan Zhao, Yifan Yang, Ruijie Zhang, Xinling Yu, and Zheng Zhang. GRZO: Group-relative zeroth-order optimization for large language model fine-tuning. *arXiv preprint arXiv:2606.02857*, 2026.

Qitao Tan, Jun Liu, Zheng Zhan, Caiwen Ding, Yanzhi Wang, Xiaolong Ma, Jaewoo Lee, Jin Lu, and Geng Yuan. Harmony in divergence: Towards fast, accurate, and memory-efficient zeroth-order LLM fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Fei Wang, Li Shen, Liang Ding, Chao Xue, Ye Liu, and Changxing Ding. Simultaneous computation and memory efficient zeroth-order optimizer for fine-tuning large language models. *arXiv preprint arXiv:2410.09823*, 2024.

Shuo Wang, Ziyu Chen, and Ming Tang. CurvZO: Adaptive curvature-guided sparse zeroth-order optimization for efficient LLM fine-tuning. In *Forty-third International Conference on Machine Learning*, 2026.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning*, pp. 681–688, 2011.

Kaiyue Wen, Tengyu Ma, and Zhiyuan Li. How sharpness-aware minimization minimizes sharpness? In *The Eleventh International Conference on Learning Representations*, 2023.

Pan Xu, Jinghui Chen, Difan Zou, and Quanquan Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

Mengxia Yu, De Wang, Qi Shan, Colorado J Reed, and Alvin Wan. The super weight in large language models. *arXiv preprint arXiv:2411.07191*, 2024.

Ziming Yu, Pan Zhou, Sike Wang, Jia Li, Mi Tian, and Hua Huang. Zeroth-order fine-tuning of LLMs in random subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4475–4485, 2025.

Ziming Yu, Shuyao Xiao, Xingyu Zhao, Sike Wang, Pan Zhou, Peiyu Zang, Xiangda Yan, Yongjie Yang, and Jia Li. SubZero+: Efficient zeroth-order LLM fine-tuning via large learning rates. *arXiv preprint arXiv:2608.15665*, 2026.

Pengyun Yue, Xuanlin Yang, Mingqing Xiao, and Zhouchen Lin. PseuZO: Pseudo-zeroth-order algorithm for training deep neural networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Kairun Zhang, Haoyu Li, Yanjun Zhao, Yifan Sun, and Huan Zhang. Learning a zeroth-order optimizer for fine-tuning LLMs. In *Forty-third International Conference on Machine Learning*, 2026.

Liang Zhang, Bingcong Li, Kiran Koshy Thekumparampil, Sewoong Oh, Michael Muehlebach, and Niao He. Zeroth-order optimization finds flat minima. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv preprint arXiv:1810.12885*, 2018.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Ying Zhang, Ömer Deniz Akyildiz, Theodoros Damoulas, and Sotirios Sabanis. Nonasymptotic estimates for stochastic gradient Langevin dynamics under local conditions in nonconvex optimization. *Applied Mathematics & Optimization*, 87(2):25, 2023.

Yanjun Zhao, Sizhe Dang, Haishan Ye, Guang Dai, Yi Qian, and Ivor Tsang. Second-order fine-tuning without pain for LLMs: A Hessian informed zeroth-order optimizer. In *The Thirteenth International Conference on Learning Representations*, 2025.

Chao Zhou, Tom Jacobs, Advait Gadhikar, and Rebekka Burkholz. Pay attention to small weights. *NeurIPS*, 2025.

APPENDIX ROADMAP

This appendix is organized as follows:

- Section A reviews related work and discusses the limitations of existing zeroth-order methods.
- Section B derives key consequences of our assumptions.
- Section C characterizes the relationship between Gaussian-smoothed objective $u_\eta(\theta)$ and the Hessian-trace regularized objective $v(\theta)$.
- Section D presents the non-asymptotic Wasserstein analysis of ZOLD.

- Section E provides implementation details and hyperparameter settings of main results, and describes the illustrative two-dimensional example of Section 5.1.
- Section F provides additional experiments complementing Section 5.3 and Section 5.4, including three-run averages with error bars and computational-efficiency comparisons across zeroth- and first-order methods, selected datasets, and model scales.

A Literature Review

A.1 Related Work

Zeroth-order optimization. Zeroth-order optimization estimates gradients using only function evaluations, typically via finite differences or random perturbations, e.g., SPSA (Spall, 1992). Classical works establish convergence guarantees for ZO methods in convex and nonconvex settings (Nesterov & Spokoiny, 2017; Shamir, 2017; Ghadimi & Lan, 2013; Liu et al., 2018; Ji et al., 2019), while lower bounds highlight an inherent dependence on dimension (Duchi et al., 2015; Jamieson et al., 2012). A major limitation of ZO methods is slow convergence in high-dimensional problems, driven by the high variance of gradient estimators, which can also adversely affect optimization stability in practice. To address this, prior work has explored variance reduction (Liu et al., 2018; Fang et al., 2018; Ji et al., 2019) and adaptive estimators (Liu et al., 2019a).

ZO for LLM fine-tuning. A growing line of work focuses on improving zeroth-order optimization for LLMs. MeZO (Malladi et al., 2023) introduces a memory-efficient implementation based on seed reuse, enabling in-place gradient estimation. Building on this, subsequent methods aim to improve efficiency, stability, and scalability: HiZOO (Zhao et al., 2025) incorporates approximate second-order information, Sparse-MeZO (Liu et al., 2025) reduces parameter updates via sparsification, while other variants explore low-rank, random subspaces, adaptive sampling, and normalization strategies (Chen et al., 2025; Yu et al., 2025; Tan et al., 2025; Behric et al., 2026; Dang et al., 2026; Wang et al., 2024). More recent approaches further explore adaptive optimization and variance reduction, including Adam-style updates without explicitly maintaining moment estimates (Cai et al., 2026), activation-guided subspaces (Lin et al., 2026), curvature-guided sparse updates (Wang et al., 2026), learned perturbation strategies (Zhang et al., 2026), and low-rank curvature-aware preconditioning (Seung et al., 2026). Other recent work investigates group-relative normalization and large-learning-rate strategies to improve the stability and efficiency of zeroth-order LLM fine-tuning (Tan et al., 2026; Yu et al., 2026). Variance-reduction and estimator-stabilization techniques have also been studied (Gautam et al., 2024; Yue et al., 2025). Despite these advances, existing approaches primarily focus on improving estimator efficiency, adapting the perturbation or update structure, or reducing the variance of zeroth-order gradient estimates. These methods often introduce additional design choices, hyperparameters, auxiliary computations, or forward evaluations, and generally do not explicitly incorporate a Langevin mechanism or target flat minima.

Flatness and Langevin-based algorithms. Flat minima are widely believed to improve generalization (Hochreiter & Schmidhuber, 1997), motivating methods such as SAM (Foret et al., 2021) and other perturbation-based sharpness minimization approaches (Chaudhari et al., 2019; Kwon et al., 2021). Recent work shows that smoothing or noise can implicitly bias optimization toward flat minima (Zhang et al., 2025; Bruno et al., 2026). However, these approaches either rely on first-order gradients (Bruno et al., 2026) or do not incorporate Langevin dynamics (Zhang et al., 2025). The theoretical study of zeroth-order Langevin schemes, particularly for non-convex optimization, remains relatively underdeveloped. Existing work can be broadly divided into zeroth-order Langevin methods for sampling and those for optimization. For sampling, Roy et al. (2022); Nakakita (2025); Sahin et al. (2026) establish non-asymptotic convergence guarantees for zeroth-order Langevin Monte Carlo (ZO-LMC) under, respectively, smoothness and functional-inequality assumptions such as a log-Sobolev inequality (Roy et al., 2022), dissipativity and regularity conditions on weak gradients (Nakakita, 2025), and smoothness together with a Poincaré inequality (Sahin et al., 2026). For optimization, Liu & Wang (2020) provide a preliminary non-asymptotic analysis of one-point zeroth-order stochastic gradient Langevin dynamics, establishing convergence toward global minimizers, while Naldi et al. (2026) derive non-asymptotic expected excess-risk bounds for zeroth-order Langevin methods under smoothness and dissipativity assumptions, with their analysis proceeding through relative entropy and Log-Sobolev inequalities. However, none of these zeroth-order Langevin results provide non-asymptotic guarantees specifically targeting global flat minima.

A.2 Challenges in Zeroth-Order LLM Fine-Tuning

Zeroth-order (ZO) optimization for LLM fine-tuning provides a memory-efficient alternative to gradient-based training by estimating gradients from function evaluations alone. Despite considerable progress, two challenges remain salient in the ZO fine-tuning of LLMs.

High estimator variance and training stability. ZO gradient estimators are known to exhibit high variance, with the variance growing with the dimension and parameter norm. This can result in noisy updates and reduced training stability, manifested as oscillatory behavior or sensitivity to hyperparameters. A growing line of work addresses these issues through improved estimator designs and variance reduction (Behric et al., 2026; Yue et al., 2025), or through stabilization strategies such as normalization and second-order approximations (Dang et al., 2026; Zhao et al., 2025). These developments yield clear empirical improvements, while typically requiring additional computational or memory resources relative to vanilla ZO updates.

Layerwise scale heterogeneity. Modern deep architectures exhibit substantial heterogeneity in gradient and parameter scales across layers. Standard ZO estimators often apply a shared perturbation scale across parameters, which may not adapt well to this substantial heterogeneity (Malladi et al., 2023; Tan et al., 2025). Recent approaches mitigate this through layerwise projections, subspace methods, or rescaling (Tan et al., 2025; Yu et al., 2025), often at the cost of additional computations or auxiliary memory.

B Discussion about the Assumptions

In this section, we show how Assumptions 4.1, 4.2, and 4.3, introduced in Section B, imply linear growth, Lipschitzness, and dissipativity of the biased estimate of the zeroth-order estimator. We further provide a bound on the second moment of the two-point estimator.

Remark B.1. *By triangle inequality and Assumption 4.2, we have, for fixed $z \in \mathbb{R}^d$,*

$$\begin{aligned}\|\nabla_\theta U(\theta+\eta z, x)\| &\le \|\nabla_\theta U(\theta+\eta z, x) - \nabla_\theta U(0,x)\| \\ &\quad + \|\nabla_\theta U(0,x) - \nabla_\theta U(0,0)\| + \|\nabla_\theta U(0,0)\| \\ &\le L_1(\|\theta\| + \eta\|z\|) + L_2\|x\| + \|\nabla U(0,0)\|.\end{aligned} \tag{19}$$

Using (7), (19), Assumption 4.1, and in Nesterov & Spokoiny (2017, Lemma 1), one obtains

$$\|\nabla u_\eta(\theta)\| \le L_1(\|\theta\| + \eta\sqrt{d}) + L_2\mathbb{E}[\|X_0\|] + \|\nabla U(0,0)\|. \tag{20}$$

By Assumptions 4.1 and 4.2, we have, for $\theta, \theta' \in \mathbb{R}^d$,

$$\|\nabla u_\eta(\theta) - \nabla u_\eta(\theta')\| \le L_1\|\theta - \theta'\|. \tag{21}$$

Being u third-order smooth as in Assumption 4.2, implies that there exists $L_3 > 0$ such that for θ, $\bar\theta \in \mathbb{R}^d$,

$$\begin{aligned}&\|u(\bar\theta) - u(\theta) - \nabla u(\theta)^T(\bar\theta - \theta) - \frac{1}{2}(\bar\theta-\theta)^T\nabla^2 u(\theta)(\bar\theta-\theta) - \frac{1}{6}\nabla^3 u(\theta)[\bar\theta-\theta]^3\| \\ &\le \frac{L_3}{24}\|\theta - \bar\theta\|^4,\end{aligned} \tag{22}$$

where $\nabla^3 u(\theta)[\bar\theta-\theta]^3 := \sum_{i,j,k\in[d]} \frac{\partial^3 u(\theta)}{\partial\theta_i\partial\theta_j\partial\theta_k}(\bar\theta_i - \theta_i)(\bar\theta_j - \theta_j)(\bar\theta_k - \theta_k)$.

Lemma B.2. *Let Assumptions 4.1 and 4.2 hold. Then*

$$\begin{aligned}\mathbb{E}_{Z,X}[\|G_\eta(\theta, Z, X)\|^2] &\le 2(4+d)^2 L_1^2\|\theta\|^2 + 4(6+d)^3\eta^2 L_1^2 \\ &\quad + 16(4+d)^2(L_2^2\mathbb{E}[\|X\|^2] + \|\nabla U(0,0)\|^2).\end{aligned} \tag{23}$$

Proof of Lemma B.2. By (5), mean value theorem and Remark B.1, we obtain for $t \in (-1, 1)$

$$\begin{aligned}\|G_\eta(\theta, z, x)\| &= \left\| \frac{U(\theta+\eta z, x) - U(\theta-\eta z, x)}{2\eta} z \right\| \\ &= \|\nabla_\theta U(\theta + t\eta z, x)\cdot z\|\|z\| \\ &\le L_1\|\theta\|\|z\|^2 + L_1\eta\|z\|^3 + (L_2\|x\| + \|\nabla U(0,0)\|)\|z\|^2.\end{aligned} \tag{24}$$

Using (24), Assumption 4.1, and (Nesterov & Spokoiny, 2017, Lemma 1) for $Z \sim \mathcal{N}(0, I_d)$, we arrive at (23). □

Remark B.3. *Let $\zeta \in (0, a/(2L_1^2))$. Using Assumptions 4.1, 4.2, and 4.3, we deduce that ∇u_η is dissipative, i.e. for any $\theta \in \mathbb{R}^d$,*

$$\langle \nabla u_\eta(\theta), \theta \rangle \geq \bar{a}\|\theta\|^2 - \bar{b}, \tag{25}$$

where

$$\begin{aligned} &\bar{a} := a - 2\zeta L_1^2 > 0, \\ &\bar{b} := ((2\zeta)^{-1} + 2\zeta L_1^2)\eta^2 d + 2\zeta L_2^2 \mathbb{E}[\|X\|^2] + 2\zeta \|\nabla U(0,0)\|^2 + b > 0. \end{aligned} \tag{26}$$

Proof. Using Assumption 4.3 and Remark B.1, and Young's inequality, one obtains, for fixed $z \in \mathbb{R}^d$

$$\begin{aligned} \langle \nabla_\theta U(\theta + \eta z, x), \theta \rangle &= \langle \nabla_\theta U(\theta + \eta z, x), \theta + \eta z \rangle - \langle \nabla_\theta U(\theta + \eta z, x), \eta z \rangle \\ &\geq a\|\theta + \eta z\|^2 - b - \zeta 2^{-1} \|\nabla_\theta U(\theta + \eta z, x)\|^2 - (2\zeta)^{-1}\eta^2 \|z\|^2 \\ &\geq (a - 2\zeta L_1^2)\|\theta\|^2 + 2a\langle \theta, \eta z \rangle + a\eta^2 \|z\|^2 - 2\zeta L_1^2 \eta^2 \|z\|^2 \\ &\quad - 2\zeta L_2^2 \|x\|^2 - 2\zeta \|\nabla U(0,0)\|^2 - b - (2\zeta)^{-1}\eta^2 \|z\|^2. \end{aligned} \tag{27}$$

Using (7), the inequality (27), and Assumption 4.1, we obtain

$$\begin{aligned} \langle \nabla u_\eta(\theta), \theta \rangle &= \mathbb{E}_{Z,X}[\nabla_\theta U(\theta + \eta Z, X)] \\ &\geq (a - 2\zeta L_1^2)\|\theta\|^2 + (a - (2\zeta)^{-1} - 2\zeta L_1^2)\eta^2 d \\ &\quad - 2\zeta L_2^2 \mathbb{E}[\|X\|^2] - 2\zeta \|\nabla U(0,0)\|^2 - b \\ &\geq \bar{a}\|\theta\|^2 - \bar{b}, \end{aligned}$$

where $\bar{a}$ and $\bar{b}$ are defined in (26). □

C Relationship between u_η and $v(\theta)$

In this section, we show how the Gaussian-smoothed version of u defined in (7) and the Hessian-trace regularized objected $v(\theta)$ given in (9) are related. For this derivation, we follow the Zhang et al. (2025, Proof of Lemma 3.5) and we show that $u_\eta(\theta)$ serves as a tractable way to access curvature information of the original objective u in (4). Let $\Pi(\theta + \eta Z)$ denote the third-order Taylor expansion of $u(\theta + \eta Z)$ around θ, i.e.

$$\Pi(\theta + \eta Z) = u(\theta) + \eta \nabla u(\theta)^\top Z + \frac{\eta^2}{2} Z^\top H(\theta) Z + \frac{\eta^3}{6} \sum_{i,j,k=1}^{d} \frac{\partial^3 u}{\partial\theta_i \partial\theta_j \partial\theta_k}(\theta)\, Z_i Z_j Z_k.$$

Taking the expectation over $Z \sim \mathcal{N}(0, I_d)$ and using (9) yields

$$\begin{aligned} \mathbb{E}[\Pi(\theta + \eta Z)] &= u(\theta) + \frac{\eta^2}{2} \mathbb{E}[Z^\top H(\theta) Z] \\ &= u(\theta) + + \frac{\eta^2}{2} \mathrm{tr}\,(H(\theta)) \\ &= v(\theta). \end{aligned} \tag{28}$$

Therefore, using (7), (28), the third-order smoothness of u (Assumption 4.2 and Remark B.1), and Nesterov & Spokoiny (2017, Lemma 1) lead to the key estimate:

$$\begin{aligned} |u_\eta(\theta) - v(\theta)| &= |\mathbb{E}[u(\theta + \eta Z)] - \mathbb{E}[\Pi(\theta + \eta Z)]| \\ &\leq \mathbb{E}[|u(\theta + \eta Z) - \Pi(\theta + \eta Z)|] \\ &\leq \frac{L_3}{24} \eta^4 (d+4)^2, \end{aligned} \tag{29}$$

with $L_3 > 0$ is the third-order smoothness constant given in (22).

D NON-ASYMPTOTIC ANALYSIS FOR ZOLD

In this section, we provide the proof of Theorem 4.4. We analyze the Wasserstein-1 error bounds of ZOLD (12) to the target flatness-biased Gibbs measure $\pi^\star_{\beta,\eta}$ by decomposing it into three components using the triangle inequality:

$$W_1\left(\mathcal{L}(\theta_k), \pi^\star_{\beta,\eta}\right) \le W_1\left(\mathcal{L}(\theta_k), \mathcal{L}(Y_k^\lambda)\right) + W_1\left(\mathcal{L}(Y_k^\lambda), \pi_\beta\right) + W_1(\pi_\beta, \pi^\star_{\beta,\eta}). \tag{30}$$

These three terms respectively quantify:

- the discrepancy between the discrete ZOLD iterates and the time-rescaled version of the continuous-time Langevin dynamics driven by the smoothed objective (32);
- the convergence of this diffusion to its invariant measure $\pi_\beta \propto \exp(-\beta u_\eta(\theta))$
- the bias between this invariant distribution and the flatness-biased target $\pi^\star_{\beta,\eta}$.

The analysis of the first two terms builds on techniques developed for stochastic gradient Langevin dynamics in Chau et al. (2021); Zhang et al. (2023), but requires nontrivial modifications to accommodate both the zeroth-order estimator and ∇u_η. The third term is controlled via a refined comparison between the potentials u_η and v, building on arguments from Bruno et al. (2026) and extending them to the present setting under third-order smoothness assumptions, while working directly in Wasserstein-1 distance.

D.1 AUXILIARY PROCESSES FOR ZOLD

To control the three terms in the decomposition (30), we introduce a collection of auxiliary continuous-time processes that connect the discrete ZOLD iterates to the invariant distribution of a Langevin diffusion driven by the smoothed objective u_η.

Langevin diffusion associated with the smoothed objective. We consider the following Langevin SDE

$$\begin{aligned} Y_0 &= \theta_0 \in \mathbb{R}^d, \\ \mathrm{d}Y_t &= -\nabla u_\eta(Y_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\,\mathrm{d}B_t, \end{aligned} \tag{31}$$

where B_t is a standard d-dimensional Brownian motion. Denote by $(\mathcal{F}_t)_{t\ge0}$ the natural filtration of $(B_t)_{t\ge0}$ and by Σ_{θ_0} the sigma-algebra generated by θ_0, and we assume that $(\mathcal{F}_t)_{t\ge0}$ is independent of Σ_{θ_0}. Under Assumption 4.2, the drift ∇u_η is Lipschitz, and therefore (31) admits a unique strong solution adapted to $(\mathcal{F}_t)_{t\ge0}$. This diffusion has invariant measure $\pi_\beta \propto \exp(-\beta u_\eta(\theta))$ and serves as a continuous-time reference process for the analysis. Notably, the appearance of the smoothed objective u_η reflects the bias induced by the zeroth-order estimator.

Time-rescaled dynamics. For a fixed step size $\lambda > 0$, define the rescaled process $Y_t^\lambda := Y_{\lambda t}$ for $t \ge 0$. A standard time-change argument yields

$$\begin{aligned} Y_0^\lambda &= \theta_0 \\ \mathrm{d}Y_t^\lambda &= -\lambda\nabla u_\eta(Y_t^\lambda)\,\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\,\mathrm{d}\widetilde{B}_t^\lambda, \end{aligned} \tag{32}$$

where $\widetilde{B}_t^\lambda := B_{\lambda t}/\sqrt{\lambda}$ is again a standard Brownian motion. This rescaling aligns the diffusion with the time scale of the discrete ZOLD iterations and is used to control the first two terms on the right-hand side of (30).

Continuous-time interpolation of ZOLD. To relate the discrete recursion (12) to the continuous-time dynamics, we introduce the interpolation $(\bar\theta_t)_{t\ge0}$ defined by

$$\begin{aligned} \bar\theta_0 &= \theta_0, \\ \mathrm{d}\bar\theta_t &= -\lambda G_\eta(\bar\theta_{\lfloor t\rfloor}, Z_{\lceil t\rceil}, X_{\lceil t\rceil})\,\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\mathrm{d}\widetilde{B}_t^\lambda, \end{aligned} \tag{33}$$

where $\lfloor r\rfloor$ and $\lceil r\rceil$ denote integer part of a positive real number r and its ceiling, respectively. By construction, the interpolation (33) coincides with the discrete iterates at integer times, i.e.,

$\mathcal{L}(\bar{\theta}_k) = \mathcal{L}(\theta_k)$ for all $k \in \mathbb{N}$. This allows us to analyze the algorithm through its continuous-time representation.

Auxiliary process for the discretization error. To localize the discretization error $W_1\left(\mathcal{L}(\theta_k), \mathcal{L}(Y_k^\lambda)\right)$ over finite time intervals, we introduce an auxiliary process $(\Phi_t^{s,\nu,\lambda})_{t\geq s}$ defined as the solution of

$$\begin{aligned} \Phi_s^{s,\nu,\lambda} &= \nu \in \mathbb{R}^d \\ \mathrm{d}\Phi_t^{s,\nu,\lambda} &= -\lambda \nabla u_\eta(\Phi_t^{s,\nu,\lambda})\,\mathrm{d}t + \sqrt{2\lambda\beta^{-1}}\,\mathrm{d}\widetilde{B}_t^\lambda. \end{aligned} \tag{34}$$

Fix $k \in \mathbb{N}$. For any $t \geq kT$, define

$$\bar{\Phi}_t^{\lambda,k} := \Phi_t^{kT,\bar{\theta}_{kT},\lambda}, \qquad \text{where} \qquad T := \lfloor 1/\lambda \rfloor. \tag{35}$$

That is, $\bar{\Phi}_t^{\lambda,k}$ evolves according to the Langevin dynamics starting from the interpolated ZOLD state at time kT. This construction enables a comparison between (32) and (33), and will be used to control the discretization error term in (30).

D.2 Preliminary Results for ZOLD Analysis

In this section, we establish some preliminary estimates on the moments of the iterates of the continuous interpolation of the ZOLD, its one-step error, the moments of the process (35), and a contraction property for the overdamped Langevin SDE (31). First, we define

$$\lambda_{\max} = \min\left\{ \frac{\min\{\bar{a}, \bar{a}^{\frac{1}{3}}\}}{16(1+L_1^2)(8+d)^4}, \frac{1}{\bar{a}}, \frac{1}{2} \right\}, \tag{36}$$

where $\bar{a}$ is given in (26).

Lemma D.1. *Let Assumptions 4.1 , 4.2, and 4.3 hold. Then, for any $\lambda \in (0, \lambda_{max})$, with λ_{max} given in (36),*

$$\sup_{t\geq 0} \mathbb{E}[\|\bar{\theta}_t\|^2] \leq C_1\left(1 + \mathbb{E}[\|\theta_0\|^2]\right), \tag{37}$$

where

$$\begin{aligned} C_1 = &\left(\frac{2\bar{b}}{\bar{a}} + 4\eta^2 \frac{(6+d)^3}{(4+d)^2} + \frac{16}{L_1^2}(L_2^2\mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2) + \frac{2d}{\beta\bar{a}} + 1 \right) \\ &\times \left(\frac{2\bar{a}^2}{L_1^2(4+d)^2} + 2 \right) \\ &+ \frac{4\bar{a}}{(4+d)^2 L_1^2}\left(\frac{\eta^2\bar{a}(6+d)^3}{(4+d)^2} + \frac{4\bar{a}(L_2^2\mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2)}{L_1^2} + \frac{d}{\beta} \right). \end{aligned} \tag{38}$$

Moreover, for any $\lambda \in (0, \lambda_{max})$, with λ_{max} given in (36),

$$\sup_{t\geq 0} \mathbb{E}\left[\|\bar{\theta}_t\|^4\right] \leq C_{1,4}\left(1 + \mathbb{E}[\|\theta_0\|^4]\right), \tag{39}$$

where

$$C_{1,4} = 1 + ((1+\bar{a}\lambda_{\max})\rho_2 + 12d^2\beta^{-2}(\lambda_{\max} + 9\bar{a}^{-1}))(\lambda_{\max} + 1/\bar{a}), \tag{40}$$

with

$$\begin{aligned} \rho_2 = &\, 4\bar{b}M^2 + 24\lambda_{max}\left(L_1^2\eta^2(6+d)^3 + 2(4+d)^2(L_2^2\mathbb{E}\left[\|X_0\|^2\right] + \|\nabla U(0,0)\|^2)\right)M^2 \\ &+ 64\lambda_{max}^2\left(L_1^3\eta^3(9+d)^{\frac{9}{2}} + 4(6+d)^3(L_2^3\mathbb{E}\left[\|X_0\|^3\right] + \|\nabla U(0,0)\|^3)\right)M \\ &+ 128\lambda_{max}^3\left(L_1^4\eta^4(12+d)^6 + 8(8+d)^4(L_2^4\mathbb{E}\left[\|X_0\|^4\right] + \|\nabla U(0,0)\|^4)\right), \end{aligned} \tag{41}$$

and

$$\begin{aligned} M = \max\Bigg\{ &\sqrt{\frac{8\bar{b}}{\bar{a}} + \frac{48\lambda_{\max}(L_1^2\eta^2(6+d)^3 + 2(4+d)^2(L_2^2\mathbb{E}\left[\|X_0\|^2\right] + \|\nabla U(0,0)\|^2))}{\bar{a}}}, \\ &\left(\frac{128\lambda_{\max}^2(L_1^3\eta^3(9+d)^{\frac{9}{2}} + 4(6+d)^3(L_2^3\mathbb{E}\left[\|X_0\|^3\right] + \|\nabla U(0,0)\|^3))}{\bar{a}} \right)^{\frac{1}{3}} \Bigg\}. \end{aligned} \tag{42}$$

Proof of Lemma D.1. Using (12), one obtains

$$\begin{aligned}\|\theta_{k+1}\|^2 = \|\theta_k\|^2 - 2\lambda\langle\theta_k, G_\eta(\theta_k, Z_{k+1}, X_{k+1})\rangle + \lambda^2\|G_\eta(\theta_k, Z_{k+1}, X_{k+1})\|^2 \\ + \frac{2\lambda}{\beta}\|\xi_{k+1}\|^2 + 2\langle\theta_k - \lambda G_\eta(\theta_k, Z_{k+1}, X_{k+1}), \sqrt{2\lambda\beta^{-1}}\xi_{k+1}\rangle.\end{aligned} \tag{43}$$

Taking the conditional expectation on θ_k on both sides of (43) and using Assumption 4.1, (6), Remark B.3, Lemma B.2, and $\lambda < \bar{a}/(2(4+d)^2L_1^2)$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\|\theta_{k+1}\|^2|\theta_k\right] \\ &= \|\theta_k\|^2 - 2\lambda\mathbb{E}\left[\langle\theta_k, G_\eta(\theta_k, Z_{k+1}, X_{k+1})\rangle|\theta_k\right] + \lambda^2\mathbb{E}\left[\|G_\eta(\theta_k, Z_{k+1}, X_{k+1})\|^2|\theta_k\right] + \frac{2\lambda d}{\beta} \\ &\le \|\theta_k\|^2 - 2\lambda\bar{a}\|\theta_k\|^2 + 2\lambda\bar{b} + \lambda^2 2(4+d)^2L_1^2\|\theta_k\|^2 + \lambda^2 4(6+d)^3\eta^2L_1^2 \\ &\quad + \lambda^2 16(4+d)^2(L_2^2\mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2) + \frac{2\lambda d}{\beta} \\ &\le (1-\lambda\bar{a})\|\theta_k\|^2 \\ &\quad + 2\lambda\left(\bar{b} + 2\bar{a}\eta^2\frac{(6+d)^3}{(4+d)^2} + \frac{8\bar{a}}{L_1^2}(L_2^2\mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2) + \frac{d}{\beta}\right).\end{aligned}$$

Taking the expectation with respect to the iterate θ_k, we arrive at

$$\begin{aligned}\mathbb{E}\left[\|\theta_{k+1}\|^2\right] &\le (1-\lambda\bar{a})^{k+1}\,\mathbb{E}[\|\theta_0\|^2] \\ &\quad + \left(\frac{2\bar{b}}{\bar{a}} + 4\eta^2\frac{(6+d)^3}{(4+d)^2} + \frac{16}{L_1^2}(L_2^2\mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2) + \frac{2d}{\beta\bar{a}}\right) \\ &\le C_0\left(1 + \mathbb{E}[\|\theta_0\|^2]\right),\end{aligned} \tag{44}$$

where

$$C_0 = \frac{2\bar{b}}{\bar{a}} + 4\eta^2\frac{(6+d)^3}{(4+d)^2} + \frac{16}{L_1^2}(L_2^2\mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2) + \frac{2d}{\beta\bar{a}} + 1. \tag{45}$$

Using (33), Hölder's inequality, and the inequality (24), one obtains

$$\begin{aligned}\|\bar{\theta}_t\|^2 &\le 2\|\bar{\theta}_t - \bar{\theta}_{\lfloor t\rfloor}\|^2 + 2\|\bar{\theta}_{\lfloor t\rfloor}\|^2 \\ &\le 4\lambda^2\left\|\int_{\lfloor t\rfloor}^t G_\eta(\bar{\theta}_{\lfloor s\rfloor}, Z_{\lceil s\rceil}, X_{\lceil s\rceil})\,\mathrm{d}s\right\|^2 + \frac{8\lambda}{\beta}\|\widetilde{B}_t^\lambda - \widetilde{B}_{\lfloor t\rfloor}^\lambda\|^2 + 2\|\bar{\theta}_{\lfloor t\rfloor}\|^2 \\ &\le 4\lambda^2(t-\lfloor t\rfloor)\int_{\lfloor t\rfloor}^t \|G_\eta(\bar{\theta}_{\lfloor s\rfloor}, Z_{\lceil s\rceil}, X_{\lceil s\rceil})\|^2\mathrm{d}s + \frac{8\lambda}{\beta}\|\widetilde{B}_t^\lambda - \widetilde{B}_{\lfloor t\rfloor}^\lambda\|^2 + 2\|\bar{\theta}_{\lfloor t\rfloor}\|^2 \\ &\le 4\lambda^2\int_{\lfloor t\rfloor}^t (2L_1^2\|Z_{\lceil s\rceil}\|^4\|\bar{\theta}_{\lfloor s\rfloor}\|^2 + 4L_1^2\eta^2\|Z_{\lceil s\rceil}\|^6\,\mathrm{d}s \\ &\quad + 32\lambda^2\int_{\lfloor t\rfloor}^t (L_2^2\mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2)\|Z_{\lceil s\rceil}\|^4)\,\mathrm{d}s \\ &\quad + \frac{8\lambda}{\beta}\|\widetilde{B}_t^\lambda - \widetilde{B}_{\lfloor t\rfloor}^\lambda\|^2 + 2\|\bar{\theta}_{\lfloor t\rfloor}\|^2.\end{aligned}$$

We note that $\lfloor s\rfloor = \lfloor t\rfloor$ for any $s \in [\lfloor t\rfloor, t]$. Taking the expectation on both sides above, using Assumption 4.1, and $\lambda < \bar{a}/(2(4+d)^2L_1^2)$, yields

$$\begin{aligned}\mathbb{E}[\|\bar{\theta}_t\|^2] &\le (8\lambda^2L_1^2(4+d)^2 + 2)\mathbb{E}[\|\bar{\theta}_{\lfloor t\rfloor}\|^2] \\ &\quad + 16\lambda^2((6+d)^3\eta^2L_1^2 + 2(4+d)^2(L_2^2\mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2)) + \frac{8\lambda d}{\beta} \\ &\le \left(\frac{2\bar{a}^2}{L_1^2(4+d)^2} + 2\right)\mathbb{E}[\|\bar{\theta}_{\lfloor t\rfloor}\|^2] \\ &\quad + \frac{4\bar{a}}{(4+d)^2L_1^2}\left(\frac{\eta^2\bar{a}(6+d)^3}{(4+d)^2} + \frac{4\bar{a}(L_2^2\mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2)}{L_1^2} + \frac{d}{\beta}\right).\end{aligned}$$

Using the fact that the law of the interpolation scheme (33) coincides with the law of the ZOLD algorithm at grid-points, i.e. $\mathcal{L}(\bar{\theta}_{\lfloor t \rfloor}) = \mathcal{L}(\theta_k)$ for $t \in [k, k+1)$, and the estimate (44), we arrive at

$$\mathbb{E}[\|\bar{\theta}_t\|^2] \leq C_1 \left(1 + \mathbb{E}[\|\theta_0\|^2]\right),$$

with C_1 given in (38). We proceed to prove (39). For $t \in (k, k+1]$, let

$$\Psi_{k,t} := \bar{\theta}_k - \lambda G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})(t-k), \tag{46}$$

and let

$$R^\lambda_{k,t} := \{2\lambda\beta^{-1}\}^{1/2}(\tilde{B}^\lambda_t - \tilde{B}^\lambda_k). \tag{47}$$

Using (33), the inequality $2z_1z_2 \leq \dot{\varepsilon} z_1^2 + \dot{\varepsilon}^{-1} z_2^2$, for $z_1, z_2 \geq 0$ and $\dot{\varepsilon} > 0$ with $z_1 = \|\Psi_{k,t}\|^2$, $z_2 = 3\|R^\lambda_{k,t}\|^2$, $\dot{\varepsilon} = \bar{a}\lambda(t-k)$, and the independence of $R^\lambda_{k,t}$ and $\bar{\theta}_k$, we obtain

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\theta}_t\|^4 \,\middle|\bar{\theta}_k\right] &= \mathbb{E}\left[\left(\|\Psi_{k,t}\|^2 + \|R^\lambda_{k,t}\|^2 + 2\left\langle \Psi_{k,t}, R^\lambda_{k,t}\right\rangle\right)^2 \middle|\bar{\theta}_k\right] \\
&= \mathbb{E}\left[\|\Psi_{k,t}\|^4 + \|R^\lambda_{k,t}\|^4 + 2\|\Psi_{k,t}\|^2\|R^\lambda_{k,t}\|^2 + 4\|\Psi_{k,t}\|^2\left\langle \Psi_{k,t}, R^\lambda_{k,t}\right\rangle \right. \\
&\qquad \left. + 4\|R^\lambda_{k,t}\|^2\left\langle \Psi_{k,t}, R^\lambda_{k,t}\right\rangle + 4\left(\left\langle \Psi_{k,t}, R^\lambda_{k,t}\right\rangle\right)^2 \middle|\bar{\theta}_k\right] \\
&\leq \mathbb{E}\left[\|\Psi_{k,t}\|^4 + \|R^\lambda_{k,t}\|^4 + 6\|\Psi_{k,t}\|^2\|R^\lambda_{k,t}\|^2 \,\middle|\bar{\theta}_k\right] \\
&\leq (1 + \bar{a}\lambda(t-k))\mathbb{E}\left[\|\Psi_{k,t}\|^4 \,\middle|\bar{\theta}^\lambda_k\right] + (1 + 9/(\bar{a}\lambda(t-k)))\mathbb{E}\left[\|R^\lambda_{k,t}\|^4\right].
\end{aligned} \tag{48}$$

By Cauchy-Schwarz inequality, we observe

$$\begin{aligned}
&\mathbb{E}\left[\|\Psi_{k,t}\|^4 \,\middle|\bar{\theta}_k\right] \\
&= \mathbb{E}\left[\left(\|\bar{\theta}_k\|^2 - 2\lambda(t-k)\left\langle \bar{\theta}_k, G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\right\rangle + \lambda^2(t-k)^2\|G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\|^2\right)^2 \middle|\bar{\theta}_k\right] \\
&= \mathbb{E}\left[\|\bar{\theta}_k\|^4 + 4\lambda^2(t-k)^2\left(\left\langle \bar{\theta}_k, G_\eta(\bar{\theta}^\lambda_k, Z_{k+1}, X_{k+1})\right\rangle\right)^2 \right. \\
&\qquad + \lambda^4(t-k)^4\|G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\|^4 - 4\lambda(t-k)\left\langle \bar{\theta}_k, G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\right\rangle\|\bar{\theta}_k\|^2 \\
&\qquad + 2\lambda^2(t-k)^2\|\bar{\theta}_k\|^2\|G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\|^2 \\
&\qquad \left. - 4\lambda^3(t-k)^3\left\langle \bar{\theta}_k, G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\right\rangle\|G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\|^2 \middle|\bar{\theta}_k\right] \\
&\leq \mathbb{E}\Bigg[6\lambda^2(t-k)^2\|\bar{\theta}_k\|^2\|G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\|^2 - 4\lambda(t-k)\left\langle \bar{\theta}_k, G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\right\rangle\|\bar{\theta}_k\|^2 \\
&\qquad - 4\lambda^3(t-k)^3\|G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\|^2\left\langle \bar{\theta}_k, G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\right\rangle \\
&\qquad + \lambda^4(t-k)^4\|G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\|^4 \Bigg|\bar{\theta}_k\Bigg] + \|\bar{\theta}_k\|^4.
\end{aligned}$$

Next, we obtain by Lemma B.2, the independence of the random variables Z_{k+1} and X_{k+1} and $\bar{\theta}_k$, that, for $q \geq 1$,

$$\begin{aligned}
\mathbb{E}\left[\|G_\eta(\bar{\theta}_k, Z_{k+1}, X_{k+1})\|^q \,\middle|\bar{\theta}_k\right] &\leq 2^{q-1}L_1^q\mathbb{E}[\|Z_{k+1}\|^{2q}]\|\bar{\theta}_k\|^q + 2^{2q-2}L_1^q\eta^q\mathbb{E}[\|Z_{k+1}\|^{3q}] \\
&\quad + 2^{3q-3}\mathbb{E}[\|Z_{k+1}\|^{2q}](L_2^q\mathbb{E}[\|X_0\|^q] + \|\nabla U(0,0)\|^q|).
\end{aligned} \tag{49}$$

Using Remark B.3, the bound (49) for $q = 2, 3, 4$, and Nesterov & Spokoiny (2017, Lemma 1), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\Psi_{k,t}\|^4 \,\middle|\bar{\theta}_k\right] \\
&\leq (1 - 4\bar{a}\lambda(t-k))\|\bar{\theta}_k\|^4 + 4\bar{b}\lambda(t-k)\|\bar{\theta}_k\|^2 \\
&\quad + 12\lambda^2(t-k)^2L_1^2(4+d)^2\|\bar{\theta}_k\|^4 + 16\lambda^3(t-k)^3L_1^3(6+d)^3\|\bar{\theta}_k\|^4 \\
&\quad + 8\lambda^4(t-k)^4L_1^4(8+d)^4\|\bar{\theta}_k\|^4 \\
&\quad + 24\lambda^2(t-k)^2\left(L_1^2\eta^2(6+d)^3 + 2(4+d)^2(L_2^2\mathbb{E}\left[\|X_0\|^2\right] + \|\nabla U(0,0)\|^2)\right)\|\bar{\theta}_k\|^2 \\
&\quad + 64\lambda^3(t-k)^3\left(L_1^3\eta^3(9+d)^{\frac{9}{2}} + 4(6+d)^3(L_2^3\mathbb{E}\left[\|X_0\|^3\right] + \|\nabla U(0,0)\|^3)\right)\|\bar{\theta}_k\| \\
&\quad + 64\lambda^4(t-k)^4\left(L_1^4\eta^4(12+d)^6 + 8(8+d)^4(L_2^4\mathbb{E}\left[\|X_0\|^4\right] + \|\nabla U(0,0)\|^4)\right),
\end{aligned}$$

which, by using $\lambda < \lambda_{\max}$, leads to

$$\begin{aligned}
&\mathbb{E}\left[\|\Psi_{k,t}\|^4 \,\middle|\bar{\theta}_k\right] \\
&\leq (1-3\bar{a}\lambda(t-k))\|\bar{\theta}_k\|^4 + 4\bar{b}\lambda(t-k)\|\bar{\theta}_k\|^2 \\
&\quad + 24\lambda^2(t-k)^2(L_1^2\eta^2(6+d)^3 + 2L_2^2\mathbb{E}\left[\|X_0\|^2\right](4+d)^2 + 2\|\nabla U(0,0)\|^2(4+d)^2)\|\bar{\theta}_k\|^2 \\
&\quad + 64\lambda^3(t-k)^3(L_1^3\eta^3(9+d)^{\frac{9}{2}} + 4L_2^3\mathbb{E}\left[\|X_0\|^3\right](6+d)^3 + 4\|\nabla U(0,0)\|^3(6+d)^3)\|\bar{\theta}_k\| \\
&\quad + 64\lambda^4(t-k)^4(L_1^4\eta^4(12+d)^6 + 8L_2^4\mathbb{E}\left[\|X_0\|^4\right](8+d)^4 + 8\|\nabla U(0,0)\|^4(8+d)^4).
\end{aligned} \tag{50}$$

We observe that, for
$\|\bar{\theta}_k\| > (8\bar{b}\bar{a}^{-1} + 48\bar{a}^{-1}\lambda_{\max}(L_1^2\eta^2(6+d)^3 + 2L_2^2\mathbb{E}\left[\|X_0\|^2\right](4+d)^2 + 2\|\nabla U(0,0)\|^2(4+d)^2))^{1/2}$,
we have

$$\begin{aligned}
&-\frac{\bar{a}\lambda(t-k)}{2}\|\bar{\theta}_k\|^4 + 4\bar{b}\lambda(t-k)\|\bar{\theta}_k\|^2 \\
&\quad + 24\lambda^2(t-k)^2\left(L_1^2\eta^2(6+d)^3 + 2(4+d)^2(L_2^2\mathbb{E}\left[\|X_0\|^2\right] + \|\nabla U(0,0)\|^2)\right)\|\bar{\theta}_k\|^2 < 0,
\end{aligned} \tag{51}$$

and for
$\|\bar{\theta}_k\| > (128\bar{a}^{-1}\lambda_{\max}^2(L_1^3\eta^3(9+d)^{\frac{9}{2}} + 4L_2^3\mathbb{E}\left[\|X_0\|^3\right](6+d)^3 + 4\|\nabla U(0,0)\|^3(6+d)^3))^{\frac{1}{3}}$,
we have

$$\begin{aligned}
&-\frac{\bar{a}\lambda(t-k)}{2}\|\bar{\theta}_k\|^4 \\
&\quad + 64\lambda^3(t-k)^3\left(L_1^3\eta^3(9+d)^{\frac{9}{2}} + 4(6+d)^3(L_2^3\mathbb{E}\left[\|X_0\|^3\right] + \|\nabla U(0,0)\|^3)\right)\|\bar{\theta}_k\| < 0.
\end{aligned} \tag{52}$$

Let M be defined as in (42). In addition, let $\mathsf{D}_{k,M} := \{\omega \in \Omega : \|\bar{\theta}_k(\omega)\| > M\}$. Using (51), (52) into (50), we have

$$\begin{aligned}
&\mathbb{E}\left[\|\Psi_{k,t}\|^4 \mathbb{1}_{\mathsf{D}_{k,M}} \,\middle|\bar{\theta}_k\right] \\
&\leq (1-2\bar{a}\lambda(t-k))\|\bar{\theta}_k\|^4 \mathbb{1}_{\mathsf{D}_{k,M}} \\
&\quad + 64\lambda^4(t-k)^4\left(L_1^4\eta^4(12+d)^6 + 8(8+d)^4(L_2^4\mathbb{E}\left[\|X_0\|^4\right] + \|\nabla U(0,0)\|^4)\right)\mathbb{1}_{\mathsf{D}_{k,M}}.
\end{aligned}$$

Analogously, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\|\Psi_{k,t}\|^4 \mathbb{1}_{\mathsf{D}_{k,M}^c} \,\middle|\bar{\theta}_k\right] \\
&\leq (1-2\bar{a}\lambda(t-k))\|\bar{\theta}_k\|^4 \mathbb{1}_{\mathsf{D}_{k,M}^c} + 4\bar{b}\lambda(t-k)M^2\mathbb{1}_{\mathsf{D}_{k,M}^c} \\
&\quad + 24\lambda^2(t-k)^2\left(L_1^2\eta^2(6+d)^3 + 2(4+d)^2(L_2^2\mathbb{E}\left[\|X_0\|^2\right] + \|\nabla U(0,0)\|^2)\right)M^2\mathbb{1}_{\mathsf{D}_{k,M}^c} \\
&\quad + 64\lambda^3(t-k)^3\left(L_1^3\eta^3(9+d)^{\frac{9}{2}} + 4(6+d)^3(L_2^3\mathbb{E}\left[\|X_0\|^3\right] + \|\nabla U(0,0)\|^3)\right)M\mathbb{1}_{\mathsf{D}_{k,M}^c} \\
&\quad + 64\lambda^4(t-k)^4\left(L_1^4\eta^4(12+d)^6 + 8(8+d)^4(L_2^4\mathbb{E}\left[\|X_0\|^4\right] + \|\nabla U(0,0)\|^4)\right)\mathbb{1}_{\mathsf{D}_{k,M}^c}.
\end{aligned}$$

Putting together the two estimates above, we arrive at

$$\mathbb{E}\left[\|\Psi_{k,t}\|^4 \,\middle|\bar{\theta}_k\right] \leq (1-2\bar{a}\lambda(t-k))\|\bar{\theta}_k\|^4 + \lambda(t-k)\rho_2, \tag{53}$$

where ρ_2 is defined in (41). Using (53) in (48), yields

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\theta}_t\|^4 \,\middle|\bar{\theta}_k\right] &\leq (1+\bar{a}\lambda(t-k))(1-2\bar{a}\lambda(t-k))\|\bar{\theta}_k\|^4 \\
&\quad + (1+\bar{a}\lambda(t-k))\lambda(t-k)\rho_2 + 12d^2\lambda^2\beta^{-2}(t-k)^2(1+9/(\bar{a}\lambda(t-k))) \\
&\leq (1-\bar{a}\lambda(t-k))\|\bar{\theta}_k\|^4 + \lambda(t-k)(1+\bar{a}\lambda_{\max})\rho_2 + 12d^2\beta^{-2}(\lambda_{\max} + 9\bar{a}^{-1}).
\end{aligned}$$

Therefore, for any $k \in \mathbb{N}, t \in (k, k+1]$, $0 < \lambda \leq \lambda_{\max}$, we have

$$\begin{aligned}
\mathbb{E}\left[\|\bar{\theta}_t\|^4\right] &\leq (1-\bar{a}\lambda(t-k))(1-\bar{a}\lambda)^k\mathbb{E}\left[\|\theta_0\|^4\right] \\
&\quad + ((1+\bar{a}\lambda_{\max})\rho_2 + 12d^2\beta^{-2}(\lambda_{\max} + 9\bar{a}^{-1}))(\lambda_{\max} + \bar{a}^{-1}),
\end{aligned}$$

which implies

$$\sup_{t\geq 0}\mathbb{E}\left[\|\bar{\theta}_t\|^4\right] \leq C_{1,4}\left(1 + \mathbb{E}[\|\theta_0\|^4]\right),$$

where $C_{1,4}$ is defined in (40). □

Next, we provide an estimate for the one-step error of the ontinuous-time interpolation of ZOLD $(\bar{\theta}_t)_{t\geq 0}$ defined in (33).

Lemma D.2. *Let Assumptions 4.1, 4.2, and 4.3 hold. Then, for any $\lambda \in (0, \lambda_{max})$, with λ_{max} given in (36),*

$$\mathbb{E}\|\bar{\theta}_t - \bar{\theta}_{\lfloor t \rfloor}\|^2 \leq C_2 \lambda (1 + \mathbb{E}[\|\theta_0\|^2]), \tag{54}$$

where

$$C_2 = 2aC_1 + \frac{4a\eta^2(6+d)^3}{(4+d)^2} + \frac{8a}{L_1^2}(L_2^2 \mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2) + \frac{4d}{\beta}, \tag{55}$$

with C_1 is given in (38).

Proof of Lemma D.2. By (33), Hölder's inequality, the inequality (24), we obtain

$$\begin{aligned}
&\|\bar{\theta}_{\lfloor t \rfloor} - \bar{\theta}_t\|^2 \\
&\leq 2 \left\| \int_{\lfloor t \rfloor}^t -\lambda G_\eta(\bar{\theta}_{\lfloor s \rfloor}, Z_{\lceil s \rceil}, X_{\lceil s \rceil}) \, \mathrm{d}s \right\|^2 + \frac{4\lambda}{\beta} \left\| \widetilde{B}^\lambda_{\lfloor t \rfloor} - \widetilde{B}^\lambda_t \right\|^2 \\
&\leq 2\lambda^2 \int_{\lfloor t \rfloor}^t (2L_1^2 \|Z_{\lceil s \rceil}\|^4 \|\bar{\theta}_{\lfloor s \rfloor}\|^2 + 4L_1^2\eta^2 \|Z_{\lceil s \rceil}\|^6 + 8\|Z_{\lceil s \rceil}\|^4 (L_2^2 \|X_{\lceil s \rceil}\|^2 + \|\nabla U(0,0)\|^2) \, \mathrm{d}s \\
&\quad + \frac{4\lambda}{\beta} \left\| \widetilde{B}^\lambda_{\lfloor t \rfloor} - \widetilde{B}^\lambda_t \right\|^2 .
\end{aligned}$$

Taking the expectation on both sides above and using Assumption 4.1, (Nesterov & Spokoiny, 2017, Lemma 1), Lemma D.1, and $\lambda < \bar{a}/(2(4+d)^2 L_1^2)$ leads to

$$\begin{aligned}
&\mathbb{E}\|\bar{\theta}_{\lfloor t \rfloor} - \bar{\theta}_t\|^2 \\
&\leq 2\lambda^2 \int_{\lfloor t \rfloor}^t 2L_1^2 (4+d)^2 C_1 \left(1 + \mathbb{E}[\|\theta_0\|^2]\right) \mathrm{d}s \\
&\quad + 2\lambda^2 \int_{\lfloor t \rfloor}^t (4L_1^2\eta^2(6+d)^3 + 8(4+d)^2 (L_2^2 \mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2)) \, \mathrm{d}s + \frac{4\lambda d}{\beta} \\
&\leq \lambda \left(2\bar{a} C_1 \left(1 + \mathbb{E}[\|\theta_0\|^2]\right) + \frac{4\bar{a}\eta^2(6+d)^3}{(4+d)^2} + \frac{8\bar{a}}{L_1^2}(L_2^2 \mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2) + \frac{4d}{\beta} \right) \\
&\leq \lambda C_2 \left(1 + \mathbb{E}[\|\theta_0\|^2]\right),
\end{aligned}$$

where C_2 is given in (55). □

We define, for each $p \geq 1$, the Lyapunov function $\widetilde{V}_p$ by $\widetilde{V}_p(\theta) := (1 + \|\theta\|^2)^{p/2}$, $\theta \in \mathbb{R}^d$. Let $\mathcal{P}_{\widetilde{V}_p}$ denote the set of $\mu \in \mathcal{P}(\mathbb{R}^d)$ satisfying $\int_{\mathbb{R}^d} \widetilde{V}_p(\theta) \, \mu(\mathrm{d}\theta) < \infty$. For $\mu, \nu \in \mathcal{P}_{\widetilde{V}_2}$, let

$$w_{1,2}(\mu, \nu) := \inf_{\varrho \in \mathcal{A}(\mu,\nu)} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} [1 \wedge \|\theta - \theta'\|](1 + \widetilde{V}_2(\theta) + \widetilde{V}_2(\theta')) \, \varrho(\mathrm{d}\theta, \mathrm{d}\theta'), \tag{56}$$

where we recall that $\mathcal{A}(\mu, \nu)$ denotes the set of probability measures ϱ on $\mathcal{B}(\mathbb{R}^{2d})$ such that its respective marginals are μ and ν. We provide the contraction property for two overdamped Langevin SDEs (31) using the functional (56).

Proposition D.3. *Let Assumptions 4.1, 4.2, and 4.3 hold. Let $(\widetilde{Y}_t)_{t \in \mathbb{R}_+}$ be the solution of (31) with initial condition $\widetilde{Y}_0 = \widetilde{\theta}_0$ which is independent of $\mathcal{F}_\infty$ and satisfies $\mathbb{E}[\|\widetilde{\theta}_0\|^2] < \infty$. Then,*

$$w_{1,2}(\mathcal{L}(Y_t), \mathcal{L}(\widetilde{Y}_t)) \leq \hat{c} e^{-\dot{c}t} w_{1,2}(\mathcal{L}(\theta_0), \mathcal{L}(\widetilde{\theta}_0)),$$

where the constants $\dot{c}$ and $\hat{c}$ are given in Eberle et al. (2019, Theorem 2.2).

Proof of Proposition D.3. It follows using Assumptions 4.1, 4.2, and 4.3, and Lemma D.6 in the Zhang et al. (2023, Proof of Proposition 4.6). □

Remark D.4. *The constants $\dot{c}$ and $\hat{c}$ in Proposition D.3 arise from the reflection coupling construction of Eberle et al. (2019). In general non-convex settings, these constants may exhibit an unfavorable (often exponential) dependence on the dimension d. This dependence is intrinsic to current contraction results for overdamped Langevin diffusions rather than specific to ZOLD. Similar behavior appears throughout the SGLD literature (Chau et al., 2021; Zhang et al., 2023; Deng et al., 2020a;b; Neufeld et al., 2025). To avoid obscuring the main arguments with cumbersome constants, we adopt the notation $\mathcal{O}(\mathrm{poly}(\dot{c}, \hat{c}))$, following Gao et al. (2022), to denote quantities that depend polynomially on these constants. While this dependence affects the quantitative rates in our bounds, it does not alter the qualitative structure of the analysis. Improving this dependence would require sharper contraction estimates beyond those currently available under the assumptions of Eberle et al. (2019, Theorem 2.2).*

Next, we establish uniform bounds for $\widetilde{V}_4(\bar{\theta}_t)$, $\widetilde{V}_2(\bar{\Phi}_t^{\lambda,k})$, and $\widetilde{V}_4(\bar{\Phi}_t^{\lambda,k})$ which will be used to control $W_1(\mathcal{L}(\bar{\Phi}_t^{\lambda,k}), \mathcal{L}(Y_t^{\lambda}))$.

Corollary D.5. *Let Assumptions 4.1, 4.2 and 4.3 hold. For any $\lambda \in (0, \lambda_{max})$, with λ_{max} given in (36), $k \in \mathbb{N}$, $t \in (k, k+1]$,*

$$\mathbb{E}[\widetilde{V}_4(\bar{\theta}_t)] \le 2C_{1,4}(1 + \mathbb{E}[\|\theta_0\|^4]) + 2,$$

where $C_{1,4}$ is given in (40).

Proof of Corollary D.5. By the definition of the Lyapunov function $\widetilde{V}_4$ and Lemma D.1, we obtain

$$\begin{aligned}\mathbb{E}[\widetilde{V}_4(\bar{\theta}_t)] &= \mathbb{E}[(1 + \|\bar{\theta}_t\|^2)^2] \\ &\le 2(1 + \mathbb{E}[\|\bar{\theta}_t\|^4]) \\ &\le 2 + 2C_{1,4}(1 + \mathbb{E}[\|\theta_0\|^4]).\end{aligned}$$

□

We establish a drift condition for the Langevin SDE (31) to obtain moment bounds on $\widetilde{V}_2(\bar{\Phi}_t^{\lambda,k})$ and $\widetilde{V}_4(\bar{\Phi}_t^{\lambda,k})$.

Lemma D.6. *Let Assumptions 4.1 and 4.3 hold. Then, for each $p \ge 2$, $\theta \in \mathbb{R}^d$,*

$$\Delta\widetilde{V}_p(\theta)\beta^{-1} - \langle \nabla u_\eta(\theta), \nabla\widetilde{V}_p(\theta)\rangle \le -\bar{\rho}(p)\widetilde{V}_p(\theta) + \tilde{\rho}(p),$$

where

$$\bar{\rho}(p) = \bar{a}p/4, \qquad \tilde{\rho}(p) = (3/4)\bar{a}p\,\widetilde{v}_p(\overline{M}_p), \tag{57}$$

with

$$\widetilde{v}_p(\overline{M}_p) = (1 + \overline{M}_p^2)^{p/2}, \quad \overline{M}_p = (1/3 + 4\bar{b}/(3\bar{a}) + 4d/(3\bar{a}\beta) + 4(p-2)/(3\bar{a}\beta))^{1/2}. \tag{58}$$

Proof of Lemma D.6. This follows by using the Langevin SDE (31) with Remark B.3 in the arguments of the Chau et al. (2021, Proof of Lemma 3.5). □

Lemma D.7. *Let Assumptions 4.1, 4.2 and 4.3 hold. For any $\lambda \in (0, \lambda_{max})$, with λ_{max} given in (36), $t \ge kT$, with $k \in \mathbb{N}$, the following inequalities hold*

$$\begin{aligned}&\mathbb{E}[\widetilde{V}_2(\bar{\Phi}_t^{\lambda,k})] \le e^{-\bar{a}\lambda t/2}C_1\mathbb{E}[\widetilde{V}_2(\theta_0)] + 3\widetilde{v}_2(\overline{M}_2) + 1, \quad \textit{and} \\ &\mathbb{E}[\widetilde{V}_4(\bar{\Phi}_t^{\lambda,k})] \le 2e^{-\bar{a}\lambda t}C_{1,4}(1 + \mathbb{E}[\|\theta_0\|^4]) + 3\widetilde{v}_4(\overline{M}_4) + 2,\end{aligned}$$

where C_1 and $C_{1,4}$ are given in (38) and (40), respectively, and $\widetilde{v}_2(\overline{M}_2)$ and $\widetilde{v}_4(\overline{M}_4)$ are defined in (58).

Proof of Lemma D.7. It follows by applying Lemma D.6, Lemma D.1, and Corollary D.5 in the Zhang et al. (2023, Proof of Lemma 4.5). □

D.3 Proof of Theorem 4.4

In this section, we provide the proof of Theorem 4.4 by deriving non-asymptotic bounds for the three terms on the right-hand side of (30). We further decompose the first error term as

$$\begin{aligned} &W_1\left(\mathcal{L}(\bar{\theta}_t), \mathcal{L}(Y_t^\lambda)\right) \\ &\leq W_1(\mathcal{L}(\bar{\theta}_t), \mathcal{L}(\bar{\Phi}_t^{\lambda,k})) + W_1(\mathcal{L}(\bar{\Phi}_t^{\lambda,k}), \mathcal{L}(Y_t^\lambda)), \quad t \in (kT, (k+1)T], \end{aligned} \tag{59}$$

and bound each term on the right-hand side separately. Before proceeding, we recall the definition of the Wasserstein-2 distance, which will be used to control the first error term in (59). For any probability measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$, the Wasserstein-2 distance is defined as

$$W_2(\mu,\nu) = \left(\inf_{\varrho \in \mathcal{A}(\mu,\nu)} \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \|x-z\|^2 \,\mathrm{d}\varrho(x,z)\right)^{\frac{1}{2}},$$

where $\mathcal{A}(\mu,\nu)$ denotes the set of probability measures on $\mathbb{R}^{2d}$ whose marginals are μ and ν.

Lemma D.8. *Let Assumptions 4.1, 4.2, and 4.3 hold. For any $\lambda \in (0, \lambda_{max})$, with λ_{max} given in (36), $t \in (kT, (k+1)T]$,*

$$W_2(\mathcal{L}(\bar{\theta}_t), \mathcal{L}(\bar{\Phi}_t^{\lambda,k})) \leq \sqrt{\lambda} C_{3,1},$$

where $C_{3,1} = \mathcal{O}(1 + d\beta^{-1/2} + \eta d^{3/2} + d^{1/2})$ and is given explicitly in (61).

Proof of Lemma D.8. By construction, the processes $(\bar{\theta}_t)_{t\geq kT}$ and $\bar{\Phi}_t^{\lambda,k}$ defined in (33) and in (35), respectively, have the same initial condition $\bar{\theta}_{kT}$ and are driven by the same Brownian motion. We denote the filtration generated by $(\bar{\theta}_{\lfloor s \rfloor})_{kT \leq s < t}$ by $\mathcal{H}_{\lfloor s \rfloor}$, and the natural filtration of $(\widetilde{B}_t^\lambda)_{t\geq 0}$ by $(\mathcal{F}_t^\lambda)_{t\geq 0}$. By using Itô's formula, the definition of the continuous-time interpolation of ZOLD algorithm (33), Assumption 4.1, (6), Young's inequality, we have

$$\begin{aligned}
&\mathbb{E}[\|\bar{\theta}_t - \bar{\Phi}_t^{\lambda,k}\|^2] \\
&= -2\lambda \int_{kT}^t \mathbb{E}[\langle \bar{\theta}_s - \bar{\Phi}_s^{\lambda,k}, G_\eta(\bar{\theta}_{\lfloor s \rfloor}, Z_{\lceil s \rceil}, X_{\lceil s \rceil}) - \nabla u_\eta(\bar{\Phi}_s^{\lambda,k})\rangle] \mathrm{d}s \\
&= -2\lambda \int_{kT}^t \mathbb{E}[\langle \bar{\theta}_s - \bar{\theta}_{\lfloor s \rfloor}, G_\eta(\bar{\theta}_{\lfloor s \rfloor}, Z_{\lceil s \rceil}, X_{\lceil s \rceil}) - \nabla u_\eta(\bar{\Phi}_s^{\lambda,k})\rangle] \mathrm{d}s \\
&\quad - 2\lambda \int_{kT}^t \mathbb{E}[\langle \bar{\theta}_{\lfloor s \rfloor} - \bar{\Phi}_s^{\lambda,k}, \mathbb{E}[G_\eta(\bar{\theta}_{\lfloor s \rfloor}, Z_{\lceil s \rceil}, X_{\lceil s \rceil}) - \nabla u_\eta(\bar{\Phi}_s^{\lambda,k}) | \mathcal{H}_{\lfloor s \rfloor} \vee \mathcal{F}_s^\lambda]\rangle] \mathrm{d}s \\
&= 2\lambda \int_{kT}^t (s - \lfloor s \rfloor) \mathbb{E}[\langle \lambda G_\eta(\bar{\theta}_{\lfloor s \rfloor}, Z_{\lceil s \rceil}, X_{\lceil s \rceil}), G_\eta(\bar{\theta}_{\lfloor s \rfloor}, Z_{\lceil s \rceil}, X_{\lceil s \rceil})\rangle] \mathrm{d}s \\
&\quad - 2\lambda \int_{kT}^t (s - \lfloor s \rfloor) \mathbb{E}[\langle \lambda G_\eta(\bar{\theta}_{\lfloor s \rfloor}, Z_{\lceil s \rceil}, X_{\lceil s \rceil}), \nabla u_\eta(\bar{\Phi}_s^{\lambda,k})\rangle] \mathrm{d}s \\
&\quad - 2\lambda \int_{kT}^t \mathbb{E}[\langle \sqrt{2\lambda\beta^{-1}}(\widetilde{B}_s^\lambda - \widetilde{B}_{\lfloor s \rfloor}^\lambda), G_\eta(\bar{\theta}_{\lfloor s \rfloor}, Z_{\lceil s \rceil}, X_{\lceil s \rceil})\rangle] \mathrm{d}s \\
&\quad - 2\lambda \int_{kT}^t \mathbb{E}[\langle \sqrt{2\lambda\beta^{-1}}(\widetilde{B}_s^\lambda - \widetilde{B}_{\lfloor s \rfloor}^\lambda), \nabla u_\eta(\bar{\Phi}_s^{\lambda,k}) - \nabla u_\eta(\bar{\Phi}_{\lfloor s \rfloor}^{\lambda,k})\rangle] \mathrm{d}s \\
&\quad - 2\lambda \int_{kT}^t \mathbb{E}[\langle \bar{\theta}_{\lfloor s \rfloor} - \bar{\Phi}_s^{\lambda,k}, \mathbb{E}[(\nabla u_\eta(\bar{\theta}_{\lfloor s \rfloor}) - \nabla u_\eta(\bar{\Phi}_s^{\lambda,k})) | \mathcal{H}_{\lfloor s \rfloor} \vee \mathcal{F}_s^\lambda]\rangle] \mathrm{d}s. \\
&\leq 3\lambda^2 \int_{kT}^t \mathbb{E}[\|G_\eta(\bar{\theta}_{\lfloor s \rfloor}, Z_{\lceil s \rceil}, X_{\lceil s \rceil})\|^2] \,\mathrm{d}s + \lambda^2 \int_{kT}^t \mathbb{E}[\|\nabla u_\eta(\bar{\Phi}_s^{\lambda,k})\|^2] \,\mathrm{d}s \\
&\quad + (2\lambda^2 d\beta^{-1}) \int_{kT}^t (s - \lfloor s \rfloor) \,\mathrm{d}s + \lambda L_1^2 \int_{kT}^t \mathbb{E}[\|\bar{\Phi}_s^{\lambda,k} - \bar{\Phi}_{\lfloor s \rfloor}^{\lambda,k}\|^2] \,\mathrm{d}s \\
&\quad + \lambda \int_{kT}^t \mathbb{E}[\|\bar{\theta}_{\lfloor s \rfloor} - \bar{\Phi}_s^{\lambda,k}\|^2 \,\mathrm{d}s + \lambda \int_{kT}^t \mathbb{E}[\|\nabla u_\eta(\bar{\theta}_{\lfloor s \rfloor}) - \nabla u_\eta(\bar{\Phi}_s^{\lambda,k})\|^2] \,\mathrm{d}s.
\end{aligned}$$

Applying Lemma B.2, Assumption 4.1, the linear growth condition (20), the definition of the auxiliary process (35), Lipschitz condition (21), $T = \lfloor 1/\lambda \rfloor$, and Lemma D.2 on the right-hand side above, we obtain

$$
\begin{aligned}
&\mathbb{E}[\|\bar{\theta}_t - \bar{\Phi}_t^{\lambda,k}\|^2] \\
&\leq 3\lambda^2 \int_{kT}^{t} 4(4+d)L_1^2 \mathbb{E}[\|\bar{\theta}_{\lfloor s \rfloor}\|^2] \, \mathrm{d}s + 3\lambda^2 \int_{kT}^{t} 8(4+d)(L_2^2 \mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2) \, \mathrm{d}s \\
&\quad + 3\lambda^2 \int_{kT}^{t} 2\eta^2(216+d^3)L_1^2 \, \mathrm{d}s + \lambda^2 \int_{kT}^{t} (2L_1^2 \mathbb{E}[\|\bar{\Phi}_s^{\lambda,k}\|^2] + 4L_1^2\eta^2 d) \, \mathrm{d}s \\
&\quad + \lambda^2 \int_{kT}^{t} 8(L_2^2 \mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2) \, \mathrm{d}s + 2\lambda d\beta^{-1} \\
&\quad + \lambda L_1^2 \int_{kT}^{t} (2\lambda^2(s - \lfloor s \rfloor) \int_{\lfloor s \rfloor}^{s} \mathbb{E}[\|\nabla u_\eta(\bar{\Phi}_r^{\lambda,k})\|^2] \, \mathrm{d}r) \\
&\quad + \lambda L_1^2 \int_{kT}^{t} 4\lambda d\beta^{-1}(s - \lfloor s \rfloor) \, \mathrm{d}s + \lambda \int_{kT}^{t} (2 + L_1^2) \mathbb{E}[\|\bar{\theta}_{\lfloor s \rfloor} - \bar{\theta}_s\|^2] \, \mathrm{d}s \\
&\quad + \lambda \int_{kT}^{t} (2 + L_1^2) \mathbb{E}[\|\bar{\theta}_s - \bar{\Phi}_s^{\lambda,k}\|^2] \, \mathrm{d}s \\
&\leq \lambda(12(4+d)L_1^2 C_1 \left(1 + \mathbb{E}[\|\theta_0\|^2]\right)) + \lambda(24(4+d)(L_2^2 \mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2)) \\
&\quad + \lambda(6L_1^2\eta^2(216+d^3)) + \lambda(L_1^2(2C_1 + 4d/\beta + 4\bar{b})(1 + \mathbb{E}[\|\theta_0\|^2]) + 4L_1^2\eta^2 d) \\
&\quad + \lambda(2d\beta^{-1} + 2L_1^2(2L_1^2(C_1 + 2d/\beta + 2\bar{b})(1 + \mathbb{E}[\|\theta_0\|^2])) \\
&\quad + \lambda(2L_1^2(4L_1^2\eta^2 d + 8(L_2^2 \mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2))) + \lambda(2 + 2L_1^2)C_2(1 + \mathbb{E}[\|\theta_0\|^2]) \\
&\quad + \lambda(2 + 2L_1^2) \int_{kT}^{t} \mathbb{E}[\|\bar{\theta}_s - \bar{\Phi}_s^{\lambda,k}\|^2] \, \mathrm{d}s.
\end{aligned} \tag{60}
$$

Using Grönwall's inequality in (60) together with $T = \lfloor 1/\lambda \rfloor$, we obtain

$$
W_2(\mathcal{L}(\bar{X}_t), \mathcal{L}(\bar{\Phi}_t^{\lambda,k})) \leq \sqrt{\mathbb{E}[\|\bar{\theta}_t - \bar{\Phi}_t^{\lambda,k}\|^2]} \leq C_{3,1}\sqrt{\lambda},
$$

where

$$
\begin{aligned}
C_{3,1} = e^{1+L_1^2} \Bigg(& 12(4+d)L_1^2 C_1 \left(1 + \mathbb{E}[\|\theta_0\|^2]\right) + 24(4+d)(L_2^2 \mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2) \\
& + 6L_1^2\eta^2(216+d^3) + L_1^2(2C_1 + 4d/\beta + 4\bar{b})(1 + \mathbb{E}[\|\theta_0\|^2]) + 4L_1^2\eta^2 d \\
& + 2d\beta^{-1} + 2L_1^2(2L_1^2(C_1 + 2d/\beta + 2\bar{b})(1 + \mathbb{E}[\|\theta_0\|^2]) \\
& + 2L_1^2(4L_1^2\eta^2 d + 8(L_2^2 \mathbb{E}[\|X_0\|^2] + \|\nabla U(0,0)\|^2)) \\
& + (2 + 2L_1^2)C_2(1 + \mathbb{E}[\|\theta_0\|^2]) \Bigg)^{1/2},
\end{aligned} \tag{61}
$$

with $C_1 = \mathcal{O}(1 + \eta^2 d^2 + d/\beta)$ defined in (38) and $C_2 = \mathcal{O}(1 + \eta^2 d^2 + d/\beta)$ defined in (55), respectively, and $\bar{b}$ given in (26). □

The following lemma provides the bound for the second term on the right-hand side of (59).

Lemma D.9. *Let Assumptions 4.1, 4.2 and 4.3 hold. For any $\lambda \in (0, \lambda_{max})$, given in (36), $t \in (kT, (k+1)T]$,*

$$
W_1(\mathcal{L}(\bar{\Phi}_t^{\lambda,k}), \mathcal{L}(Y_t^\lambda)) \leq \sqrt{\lambda} C_{3,2},
$$

where

$$
C_{3,2} = \frac{\hat{c}}{1 - \exp(-\dot{c})} \left(C_{3,1}^2 + 3 \left[5 + 4C_{1,4}(1 + \mathbb{E}[\|\theta_0\|^4]) + 3\widetilde{v}_4(\overline{M}_4) \right] \right), \tag{62}
$$

with $C_{3,1}$ is given in (61), $C_{1,4}$ is given in (40), $\widetilde{v}_4(\overline{M}_4)$ is defined in (58), and the constants $\dot{c}$ and $\hat{c}$ are given in Eberle et al. (2019, Theorem 2.2).

Proof of Lemma D.9. This follows using Proposition D.3, Lemma D.8, Corollary D.5 and Lemma D.7 in the Zhang et al. (2023, Proof of Corollary 4.9). □

Next, we derive the non-asymptotic bound for the second term on the right-hand side of (30).

Theorem D.10. *Let Assumptions 4.1, 4.2, and 4.3 hold. Then, there exist constants $\dot{c}, \hat{c} > 0$ such that, for every $\beta > 0$, for $0 < \lambda < \lambda_{max}$ given in (36), any $t \in (kT, (k+1)T]$, and $k \in \mathbb{N}$,*

$$W_1(\mathcal{L}(Y_t^\lambda), \pi_\beta) \leq \hat{c} e^{-\dot{c}\lambda t} \left(1 + \mathbb{E}[\widetilde{V}_2(\theta_0)] + \int_{\mathbb{R}^d} \widetilde{V}_2(\theta)\pi_\beta(\mathrm{d}\theta)\right),$$

where the constants $\dot{c}$ and $\hat{c}$ are given in Eberle et al. (2019, Theorem 2.2)..

Proof of Theorem D.10. This follows using $W_1(\mu, \nu) \leq w_{1,2}(\mu, \nu)$, Proposition D.3, and

$$w_{1,2}(\mathcal{L}(\theta_0), \pi_\beta) \leq \left(1 + \mathbb{E}[\widetilde{V}_2(\theta_0)] + \int_{\mathbb{R}^d} \widetilde{V}_2(\theta)\pi_\beta(\mathrm{d}\theta)\right).$$

□

The last error bound on the right-hand side of (30) is derived using the following two auxiliary results.

Lemma D.11 (Lemma 3.5 in Zhang et al. (2025))**.** *Let Assumption 4.2 hold. Then, it holds that*

$$|u_\eta(\theta) - v(\theta)| \leq \frac{L_3}{24}\eta^4(d+4)^2, \qquad \forall\, \theta \in \mathbb{R}^d.$$

The proof of Lemma D.11 is given in Zhang et al. (2025, Appendix B.2) and is reproduced in Appendix C for completeness.

Lemma D.12. *Let Assumptions 4.1 and 4.3 hold. Then, the following set*

$$\mathcal{S} := \left\{\theta \in \mathbb{R}^d : \|\theta\| \leq \sqrt{\frac{\bar{b}}{\bar{a}}}\right\}, \tag{63}$$

contains all the minimizers of $u_\eta(\theta)$, where $\bar{a}$ and $\bar{b}$ are given in (26).

Proof of Lemma D.12. It follows using Remark B.3 in the proof of (Bruno et al., 2026, Lemma C.2). □

Proposition D.13. *Let Assumptions 4.1, 4.2, and 4.3 hold. Then,*

$$W_1(\pi_\beta, \pi^\star_{\beta,\eta}) \leq C_4,$$

where $C_4 = \widetilde{\mathcal{O}}(\eta^2 d^{\frac{3}{2}} + \sqrt{\beta}\eta^2 d + \sqrt{\beta}\eta^3 d^{\frac{3}{2}} + \sqrt{\beta}\eta^4 d^2)$ and its explicit expression is given in (68).

Proof of Proposition D.13. Let $Z_\beta := \int_{\mathbb{R}^d} e^{-\beta u_\eta(\theta)}\, \mathrm{d}\theta$ and $Z_{\beta,\eta} := \int_{\mathbb{R}^d} e^{-\beta v(\theta)}\, \mathrm{d}\theta$. Let $\mathrm{KL}(\pi_\beta || \pi^\star_{\beta,\eta})$ be the Kullback-Leibler divergence between π_β and $\pi^\star_{\beta,\eta}$. Using Lemma D.11, we have

$$\begin{aligned}
\mathrm{KL}(\pi_\beta||\pi^\star_{\beta,\eta}) &= \int_{\mathbb{R}^d} \log\left(\frac{Z_\beta^{-1} e^{-\beta u_\eta(\theta)}}{Z_{\beta,\eta}^{-1} e^{-\beta v(\theta)}}\right)\pi_\beta(\mathrm{d}\theta) \\
&= \log\left(\frac{\int_{\mathbb{R}^d} e^{-\beta u_\eta(\theta) + \beta(u_\eta(\theta) - v(\theta))}\, \mathrm{d}\theta}{\int_{\mathbb{R}^d} e^{-\beta u_\eta(\theta)}\, \mathrm{d}\theta}\right) - \beta \int_{\mathbb{R}^d} (u_\eta(\theta) - v(\theta))\pi_\beta(\mathrm{d}\theta) \qquad (64) \\
&\leq \frac{L_3}{12}\beta\eta^4(d+4)^2.
\end{aligned}$$

We denote the complementary set of $\mathcal{S}$ in Lemma D.12 by $\mathcal{S}^c$. For $\theta \in \mathcal{S}^c$ and $\widetilde{c} \in (0,1)$, we have, by Remark B.3,

$$\begin{aligned} u_\eta(\theta) &= u_\eta(\widetilde{c}\theta) + \int_{\widetilde{c}}^1 \langle \theta, \nabla u_\eta(t\theta) \rangle \, \mathrm{d}t \\ &\geq u_\eta(\theta^\star_{u_\eta}) + \int_{\widetilde{c}}^1 t^{-1} \langle t\theta, \nabla u_\eta(t\theta) \rangle \, \mathrm{d}t \\ &\geq u_\eta(\theta^\star_{u_\eta}) + \int_{\widetilde{c}}^1 t^{-1} (\bar{a}\|t\theta\|^2 - \bar{b}) \, \mathrm{d}t \\ &\geq \frac{\bar{a}(1-\widetilde{c}^2)}{2}\|\theta\|^2 + \bar{b}\log\widetilde{c} + u_\eta(\theta^\star_{u_\eta}) \\ &= \bar{c}\|\theta\|^2 + \bar{b}\log\widetilde{c} + u_\eta(\theta^\star_{u_\eta}), \end{aligned} \tag{65}$$

where $\bar{c} := \frac{\bar{a}(1-\widetilde{c}^2)}{2} > 0$ and $\theta^\star_{u_\eta}$ is a minimizer of u_η. Let $\gamma \in (0, 1 \wedge \beta\bar{c})$ and $c_{v_\beta} := \min_{\theta\in\mathcal{S}} e^{-\beta v(\theta)}$. Using Assumption 4.2, (65), and Lemma D.12, we have

$$\begin{aligned} &\int_{\mathbb{R}^d} e^{(\beta\bar{c}-\gamma)\|\theta\|^2} \pi^\star_{\beta,\eta}(\mathrm{d}\theta) \\ &= \frac{\int_{\mathbb{R}^d} e^{(\beta\bar{c}-\gamma)\|\theta\|^2 - \beta v(\theta)} \, \mathrm{d}\theta}{\int_{\mathbb{R}^d} e^{-\beta v(\theta)} \, \mathrm{d}\theta} \\ &\leq \frac{\int_{\mathcal{S}} e^{(\beta\bar{c}-\gamma)\|\theta\|^2 - \beta v(\theta)} \, \mathrm{d}\theta}{\int_{\mathcal{S}} e^{-\beta v(\theta)} \, \mathrm{d}\theta} + \frac{\int_{\mathcal{S}^c} e^{(\beta\bar{c}-\gamma)\|\theta\|^2 - \beta u_\eta(\theta) + \beta(u_\eta(\theta)-v(\theta))} \, \mathrm{d}\theta}{\int_{\mathcal{S}} e^{-\beta v(\theta)} \, \mathrm{d}\theta} \\ &\leq e^{\frac{(\beta\bar{c}-\gamma)\bar{b}}{\bar{a}}} + e^{\frac{L_3}{24}\beta\eta^4(d+4)^2} \frac{\int_{\mathcal{S}^c} e^{(\beta\bar{c}-\gamma)\|\theta\|^2 - \beta u_\eta(\theta)} \, \mathrm{d}\theta}{\int_{\mathcal{S}} e^{-\beta v(\theta)} \, \mathrm{d}\theta} \\ &\leq e^{\frac{(\beta\bar{c}-\gamma)\bar{b}}{\bar{a}}} + e^{\frac{L_3}{24}\beta\eta^4(d+4)^2} \frac{e^{-\beta\bar{b}\log\widetilde{c} - \beta u_\eta(\theta^\star_{u_\eta})} \int_{\mathcal{S}^c} e^{-\|\theta\|^2\gamma} \, \mathrm{d}\theta}{c_{v_\beta}(\frac{\bar{b}}{\bar{a}})^{\frac{d}{2}} \frac{\pi^{d/2}}{\Gamma(d/2+1)}} \\ &\leq e^{\frac{(\beta\bar{c}-\gamma)\bar{b}}{\bar{a}}} + \frac{e^{-\beta\bar{b}\log\widetilde{c} - \beta u_\eta(\theta^\star_{u_\eta}) + \frac{L_3}{24}\beta\eta^4(d+4)^2} \int_{\mathcal{S}^c} e^{-\|\theta\|^2\gamma} \, \mathrm{d}\theta}{c_{v_\beta}(\frac{\bar{b}}{\bar{a}})^{\frac{d}{2}} \frac{\pi^{d/2}}{\Gamma(d/2+1)}} \\ &\leq e^{\frac{(\beta\bar{c}-\gamma)\bar{b}}{\bar{a}}} + \frac{e^{-\beta\bar{b}\log\widetilde{c} - \beta u_\eta(\theta^\star_{u_\eta}) + \frac{L_3}{24}\beta\eta^4(d+4)^2} \bar{a}^{\frac{d}{2}}\Gamma(\frac{d}{2}+1)}{c_{v_\beta}\gamma^{\frac{d}{2}}\bar{b}^{\frac{d}{2}}} < \infty. \end{aligned} \tag{66}$$

Applying (Bolley & Villani, 2005, Corollary 2.4), using (66), (64), and the inequality $\log(x+y) \leq \log(2) + \log(x) \vee \log(y)$ for $x, y > 0$, we have

$$\begin{aligned} &W_1(\pi_\beta, \pi^\star_{\beta,\eta}) \\ &\leq \left[\frac{2}{\beta\bar{c}-\gamma}\left(1 + \log\left(\int_{\mathbb{R}^d} e^{(\beta\bar{c}-\gamma)\|\theta\|^2}\pi^\star_{\beta,\eta}(\mathrm{d}\theta)\right)\right)\right]^{\frac{1}{2}} [\mathrm{KL}(\pi_\beta||\pi^\star_{\beta,\eta})]^{\frac{1}{2}} \\ &\leq \left[\frac{2}{\beta\bar{c}-\gamma} + \frac{2\log(2)}{\beta\bar{c}-\gamma}\right. \\ &\quad \left. + \frac{2}{\beta\bar{c}-\gamma}\left(\frac{\bar{b}(\beta\bar{c}-\gamma)}{\bar{a}} \vee \log\left(\frac{e^{-\beta\bar{b}\log\widetilde{c} - \beta u_\eta(\theta^\star_{u_\eta})} e^{\frac{L_3}{24}\beta\eta^4(d+4)^2} \bar{a}^{\frac{d}{2}}\Gamma(\frac{d}{2}+1)}{c_{v_\beta}\gamma^{\frac{d}{2}}\bar{b}^{\frac{d}{2}}}\right)\right)\right]^{\frac{1}{2}} \\ &\quad \times \left[\frac{L_3}{12}\beta\eta^4(d+4)^2\right]^{\frac{1}{2}}. \end{aligned} \tag{67}$$

By Stirling's formula, we have $\log(\Gamma(d/2+1)) = \mathcal{O}(d\log d)$ and by (26), we have $\bar{b}$ is $\mathcal{O}(1+d\eta^2)$ and $\bar{a}$ is $\mathcal{O}(1)$. Therefore,

$$\begin{aligned} & W_1(\pi_\beta, \pi^\star_{\beta,\eta}) \\ & \leq \Bigg[\frac{2}{\beta\bar{c}-\gamma} + \frac{2\log(2)}{\beta\bar{c}-\gamma} \\ & \quad + 2\Bigg(\frac{\bar{b}}{\bar{a}} \vee \frac{\beta\bar{b}\log\frac{1}{\bar{c}}}{\beta\bar{c}-\gamma} - \frac{\beta u_\eta(\theta^\star_{u_\eta})}{\beta\bar{c}-\gamma} + \frac{L_3\beta\eta^4(d+4)^2}{24(\beta\bar{c}-\gamma)} + \frac{d\log(\bar{a})}{2(\beta\bar{c}-\gamma)} \\ & \qquad + \frac{\log(\Gamma(d/2+1))}{\beta\bar{c}-\gamma} - \frac{\log(c_{v_\beta})}{\beta\bar{c}-\gamma} - \frac{d}{2(\beta\bar{c}-\gamma)}\log\left(\gamma\bar{b}\right) \Bigg) \Bigg]^{\frac{1}{2}} \\ & \quad \times \left[(12)^{-\frac{1}{2}} (L_3\beta)^{\frac{1}{2}} \eta^2 (d+4) \right] =: C_4. \end{aligned} \tag{68}$$

Therefore, $C_4 = \widetilde{\mathcal{O}}(\eta^2 d^{\frac{3}{2}} + \sqrt{\beta}\eta^2 d + \sqrt{\beta}\eta^3 d^{\frac{3}{2}} + \sqrt{\beta}\eta^4 d^2)$. □

Lastly, we present the proof of Theorem 4.4 in which we establish the non-asymptotic W_1 error bounds of ZOLD to the flatness-biased target measure $\pi^*_{\beta,\eta} \propto \exp(-\beta v(\theta))$.

Proof of Theorem 4.4. Fix an arbitrary integer $k \geq 1$, and choose $j \in \mathbb{N}_0$ such that $k \in (jT, (j+1)T]$. Using $\lambda T > 1/2$, $\eta = \beta^{-\frac{1+\tau}{4}}$, Lemma D.8, Lemma D.9, Theorem D.10, and Proposition D.13 in (30), we obtain

$$\begin{aligned} W_1(\mathcal{L}(\theta_k), \pi^\star_{\beta,\eta}) &\leq W_1(\mathcal{L}(\bar{\theta}_k), \mathcal{L}(\bar{\Phi}^{\lambda,j}_k)) + W_1(\mathcal{L}(\bar{\Phi}^{\lambda,j}_k), \mathcal{L}(Y^\lambda_k)) \\ &\quad + W_1(\mathcal{L}(Y^\lambda_k), \pi_\beta) + W_1(\pi_\beta, \pi^\star_{\beta,\eta}) \\ &\leq W_2(\mathcal{L}(\bar{\theta}_k), \mathcal{L}(\bar{\Phi}^{\lambda,j}_k)) + W_1(\mathcal{L}(\bar{\Phi}^{\lambda,j}_k), \mathcal{L}(Y^\lambda_k)) \\ &\quad + W_1(\mathcal{L}(Y^\lambda_k), \pi_\beta) + W_1(\pi_\beta, \pi^\star_{\beta,\eta}) \\ &\leq C_{3,1}\sqrt{\lambda} + W_1(\mathcal{L}(\bar{\Phi}^{\lambda,j}_k), \mathcal{L}(Y^\lambda_k)) + W_1(\mathcal{L}(Y^\lambda_k), \pi_\beta) + C_4 \\ &\leq (C_{3,1} + C_{3,2})\sqrt{\lambda} + D_1 e^{-\dot{c}\lambda k}\left(1 + \mathbb{E}[\|\theta_0\|^4]\right) + C_4, \end{aligned} \tag{69}$$

where $C_{3,1}$, $C_{3,2}$, and C_4 take into account the coupling $\eta = \beta^{-\frac{1+\tau}{4}}$, and

$$D_1 = \hat{c}\left(1 + \int_{\mathbb{R}^d} \widetilde{V}_2(\theta)\pi_\beta(\mathrm{d}\theta)\right). \tag{70}$$

In addition, for any $\epsilon > 0$, if $\beta > 0$ is sufficiently large, and we choose

$$\lambda \leq \mathcal{O}\left(\frac{\epsilon^2}{d^3}\dot{c}^2\beta^{(1+\tau)/2}\right) \wedge \lambda_{\max}, \tag{71}$$

and

$$k \geq \widetilde{\mathcal{O}}\left(\frac{d^3}{\epsilon^2\beta^{(1+\tau)/2}}\mathrm{poly}\left(\frac{1}{\dot{c}}\right)\right), \tag{72}$$

then $W_1(\mathcal{L}(\theta_k), \pi^\star_{\beta,\eta}) \leq \epsilon$. □

E Experiment setup

E.1 ZOLD Hyperparameter Setting

Table 5 summarizes the hyperparameter settings used for ZOLD and LN-ZOLD. For all experiments, we set the weight decay to 0 and use β scheduling with $\beta_0 = 10^8$, $\beta_K = 10^{12}$, and $\tau = 0.5$.

Table 5: Hyperparameter settings for ZOLD and LN-ZOLD.

Experiment	Hyperparameters	Values
OPT / LLaMA3	Batch size	16
	Learning rate	{1e−7, 2e−7}
	Iterations	20k
	Weight decay	0
	τ	0.5
	β scheduling	$\beta_0 = 10^8,\ \beta_K = 10^{12}$
Phi-2	Batch size	16
	Learning rate	{1e−6, 2e−6}
	Iterations	20k
	Weight decay	0
	τ	0.5
	β scheduling	$\beta_0 = 10^8,\ \beta_K = 10^{12}$

E.2 Details of the Illustrative Two-Dimensional Example

We describe the synthetic objective used in Figure 1. Let $\theta = (\theta^{(1)}, \theta^{(2)}) \in \mathbb{R}^2$. The objective is constructed as a sum of three attractive Gaussian wells, one repulsive Gaussian barrier, and a mild quadratic confining term:

$$\begin{aligned} U(\theta) = c\|\theta\|_2^2 - a_s \exp\left(-\frac{1}{2}\left\|\frac{\theta - m_s}{\sigma_s}\right\|_2^2\right) - a_\ell \exp\left(-\frac{1}{2}\left\|\frac{\theta - m_\ell}{\sigma_\ell}\right\|_2^2\right) \\ - a_g \exp\left(-\frac{1}{2}\left\|\frac{\theta - m_g}{\sigma_g}\right\|_2^2\right) + a_b \exp\left(-\frac{1}{2}\left\|\frac{\theta - m_b}{\sigma_b}\right\|_2^2\right) + 2.3. \end{aligned} \tag{73}$$

Here division by σ is component-wise. The first Gaussian well forms the sharp minimum, the second forms the nearby local flat minimum, and the third forms the lower global flat minimum. The positive Gaussian term creates an energy barrier between the two flat basins.

The parameters are

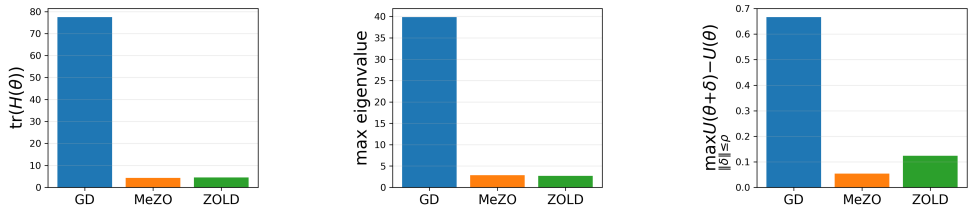
$$\begin{gathered} c = 0.05, \qquad m_s = (1.2, 0.95), \quad \sigma_s = (0.22, 0.22), \quad a_s = 2.0, \\ m_\ell = (-2.0, 1.0), \quad \sigma_\ell = (1.2, 0.8), \quad a_\ell = 2.0, \\ m_g = (-2.2, -1.7), \quad \sigma_g = (1.25, 0.95), \quad a_g = 2.8, \\ m_b = (-2.0, -0.35), \quad \sigma_b = (0.55, 0.35), \quad a_b = 1.5. \end{gathered}$$

All methods are initialized at $\theta_0 = (0.7, 0.85)$ and run with learning rate $\lambda = 0.05$ for $K = 200$ iterations. For MeZO, we use a fixed smoothing radius $\eta = 0.9$. For ZOLD, we set $\tau = 0.5$, $\beta_0 = 5 \times 10^{-1}$, and $\beta_K = 5 \times 10^5$.

In Figure 4, we further compare the sharpness of the final solutions obtained by each method using the Hessian trace, the maximum eigenvalue, and the zeroth-order sharpness.

F Additional Experiments

To further support the results in Section 5, we provide a comprehensive evaluation of ZOLD and LN-ZOLD along three axes: robustness across random seeds and model scales, comparison with first-order fine-tuning, and end-to-end computational cost. We evaluate OPT-125M, OPT-2.7B, and OPT-13B on SST-2 and WSC, repeating each experiment over three random seeds and reporting the mean and sample standard deviation of accuracy. In addition to the zeroth-order baselines (MeZO, HiZOO-L, and FZOO), we include first-order full fine-tuning (FO) and LoRA fine-tuning (FO-LoRA). Since our work focuses on memory-constrained settings in which backpropagation is infeasible, these first-order methods are not direct competitors but serve as references to quantify the accuracy–memory trade-off of zeroth-order fine-tuning. For computational cost, we report elapsed time to the best-accuracy checkpoint, peak GPU memory, and total FLOPs, each averaged over the three runs.



(a) Hessian trace (b) Maximum eigenvalue (c) Zeroth-order sharpness

Figure 4: Sharpness measures evaluated at the final iterate θ_K for each optimization algorithm.

Table 6: Performance and computational cost comparison of optimization methods using OPT-125M, OPT-2.7B, and OPT-13B on the SST-2 and WSC datasets. Accuracy is reported as the mean and sample standard deviation over three random seeds, while elapsed time, peak GPU memory, and Total FLOPs are averaged over the same runs. The best and second-best accuracies among zeroth-order methods are highlighted in **bold** and underlined, respectively.

Model	Method	SST-2				WSC			
		Accuracy (%)	Elapsed Time (min)	Peak GPU Memory (GB)	Total FLOPs	Accuracy (%)	Elapsed Time (min)	Peak GPU Memory (GB)	Total FLOPs
OPT-125M	MeZO	84.17 ± 0.70	3.88	0.868	2.72×10^{15}	62.82 ± 0.56	2.50	1.326	2.99×10^{15}
	HiZOO-L	84.25 ± 0.74	9.29	0.904	4.08×10^{15}	64.42 ± 1.67	7.46	1.328	5.50×10^{15}
	FZOO	$\mathbf{86.12 \pm 0.53}$	41.97	0.869	1.87×10^{16}	63.78 ± 1.47	27.57	1.327	2.19×10^{16}
	ZOLD	$\underline{85.63 \pm 0.37}$	9.84	0.868	6.06×10^{15}	$\mathbf{65.06 \pm 1.11}$	3.21	1.326	3.59×10^{15}
	LN-ZOLD	85.36 ± 0.18	15.15	0.868	6.07×10^{15}	$\mathbf{65.06 \pm 1.11}$	5.95	1.326	4.42×10^{15}
	FO	86.77 ± 0.13	2.78	1.385	4.37×10^{15}	63.78 ± 1.47	0.84	2.410	2.29×10^{15}
	FO-LoRA	87.46 ± 0.43	3.71	1.310	3.80×10^{15}	61.86 ± 2.00	1.48	2.275	2.91×10^{15}
OPT-2.7B	MeZO	92.74 ± 0.43	148.89	6.092	1.86×10^{17}	63.78 ± 0.56	111.54	7.147	2.78×10^{17}
	HiZOO-L	92.70 ± 0.37	139.15	7.115	2.80×10^{17}	64.10 ± 0.56	121.54	7.373	3.88×10^{17}
	FZOO	$\mathbf{94.46 \pm 0.07}$	245.11	6.092	5.70×10^{17}	$\underline{64.74 \pm 1.11}$	187.98	7.147	6.89×10^{17}
	ZOLD	93.92 ± 0.23	44.57	6.092	1.38×10^{17}	$\mathbf{65.38 \pm 2.54}$	54.68	7.146	2.55×10^{17}
	LN-ZOLD	$\underline{94.11 \pm 0.37}$	95.73	6.092	2.06×10^{17}	64.42 ± 0.96	81.64	7.146	2.88×10^{17}
	FO	93.50 ± 0.53	19.65	41.940	7.27×10^{16}	60.90 ± 4.54	16.41	50.813	8.40×10^{16}
	FO-LoRA	95.18 ± 0.53	6.03	21.247	3.54×10^{16}	62.18 ± 2.00	34.04	29.213	2.29×10^{17}
OPT-13B	MeZO	91.36 ± 0.35	318.35	25.993	1.30×10^{18}	64.10 ± 1.47	95.86	27.864	5.91×10^{17}
	HiZOO-L	91.86 ± 0.30	249.69	28.276	4.87×10^{18}	$\mathbf{64.74 \pm 1.47}$	372.92	28.425	1.52×10^{18}
	FZOO	$\mathbf{93.85 \pm 0.66}$	1936.99	25.993	4.96×10^{18}	63.78 ± 3.64	774.99	27.864	3.36×10^{18}
	ZOLD	93.16 ± 0.29	265.63	25.993	9.50×10^{17}	63.78 ± 1.11	119.52	27.864	6.97×10^{17}
	LN-ZOLD	$\underline{93.31 \pm 0.24}$	396.75	25.993	1.17×10^{18}	$\mathbf{64.74 \pm 0.56}$	201.35	27.864	9.13×10^{17}
	FO	92.13 ± 0.52	18.68	194.938	1.27×10^{17}	61.22 ± 4.00	38.97	204.648	3.31×10^{17}
	FO-LoRA	95.64 ± 0.53	65.65	85.683	5.83×10^{17}	63.14 ± 2.42	94.85	101.882	9.33×10^{17}

We use the same experimental settings as in the main experiments, except for the hyperparameter tuning described below. Each additional experiment is repeated over three same random seeds, and we report the mean accuracy and sample standard deviation. Elapsed time, peak GPU memory, and Total FLOPs are averaged over the same three runs. For MeZO, HiZOO-L, and FZOO, we follow the hyperparameter tuning protocols used in their respective original works. We tune the learning rate over {`1e-5, 5e-5, 8e-5`} for first-order full fine-tuning (FO) and {`1e-4, 5e-4, 8e-4`} for FO-LoRA. For ZOLD and LN-ZOLD, we additionally search over {`1e-7, 2e-7, 5e-7, 1e-6`} to account for the increased model scale. The best and second-best accuracies among zeroth-order methods are highlighted in bold and underlined, respectively.

Table 6 compares accuracy and computational cost across OPT model scales on SST-2 and WSC. On SST-2, ZOLD and LN-ZOLD achieve accuracy comparable to FZOO at a substantially lower computational cost. For example, on OPT-13B, ZOLD attains 93.16% accuracy compared with 93.85% for FZOO, while running $7.3\times$ faster with $5.2\times$ fewer total FLOPs. On WSC, our methods achieve the highest accuracy among ZO methods on OPT-125M and OPT-2.7B, and LN-ZOLD matches the best ZO accuracy on OPT-13B.